DOI: 10.2478/9788366675513-fm

Proceedings of the 2023 Aarhus International Conference on Voice Studies

EDITORS

Míša Hejná, *Jens Kjeldgaard-Christiansen*, *Mark Eaton*, *Mathias Clasen*, *Zac Boyd*, *Oliver Niebuhr*

MANAGING EDITORS

Míša Hejná, Oliver Niebuhr

ORGANISERS

Míša Hejná, *Mark Eaton*, *Jens Kjeldgaard-Christiansen*, *Mathias Clasen*, *Zac Boyd*

PUBLISHER

https://sciendo.com/

SCIENTIFIC COMMITTEE

Yonatan Ungermann Goldshtein
Katariina Harjunpää
Lisa Jansen
Stina Hasse Jørgensen
Monika Konert-Panek
Philippe Monneret
Paola Torres Nuñez del Prado
Birgitte Stougaard Pedersen
George Walkden
Valentin Werner
Michael Westphal
Hannah White
Jessica Wormald
and six anonymous reviewers

GRAPHIC DESIGN

Míša Hejná

(cover: “Inútil Paisagem performed by Esperanza Spalding and Gretchen Parlato”)

ISBN

978-83-66675-51-3

DOI: https://doi.org/10.2478/9788366675513

DOI: 10.2478/9788366675513-toc

Proceedings of the 2023 Aarhus International Conference on Voice Studies

TABLE OF CONTENTS

Introduction

Why interdisciplinary approaches to voice studies? pp. 1–5
—*Misa Hejna*, *Jens Kjeldgaard-Christiansen*, *and Mark Eaton*

Sociolinguistics

Intersections of sexual orientation, bilingualism, and /s/ variation in French and German men's speech pp. 6–15
—*Zac Boyd*

Assessing creaky voice quality for forensic purposes pp. 16–26
—*Katharina Klug*, *Christin Kirchhübel*, *Paul Foulkes*, *Almut Braun*, *and Peter French*

Age and gender variation in Scottish voice quality beyond creak: a multi-measure approach pp. 27–34
—*Joe Pearce*

Media and Popular Culture Studies

"Vocal interactions": the use of voice in spoken word duos pp. 35–40
—*Rebecka Dürr*

The 'Rebellious Girl Who Doesn't Try Too Hard': indexicality and independence in the music of Lily Allen pp. 41–47
—*Josephine Iben Kjeldbjerg*

The good, the bad, and the funny Russian: linguistic variation in telecinema pp. 48–55
—*Freja Hovgaard Knudsen*

Linguistic Typology

Voice quality changes, issue arises: a new typological profile of phonation in Zhangzhou pp. 56–66
—*Yishan Huang and Michael Proctor*

Voice your opinion! Does modality affect the acceptability of extraction from adjunct clauses in English? pp. 67–71
—*Anne Mette Nyvad*, *Ken Ramshøj Christensen*, *and Julie Maria Rohde*

For a systematic semantic classification of imitative prosody pp. 72–78
—*Antoine Tholly*

Digital Studies

"A human way to talk": poetry and synthetic voice pp. 79–83
—*Vadim Keylin*

The conversational (human?) voice of artificial intelligence pp. 84–91
—*Jeanette Landgrebe*

Crying is the name of the game: crying as a resource in video-mediated social-psychiatric recovery meetings pp. 92–101
—*Jeanette Landgrebe*

DOI: 10.2478/9788366675513-001

Why interdisciplinary approaches to voice studies?

Míša Hejná[1], *Jens Kjeldgaard-Christiansen*[1], *Mark Eaton*[1]

[1]Department of English, Aarhus University, Denmark
misa.hejna@cc.au.dk, jkc@cc.au.dk, engme@cc.au.dk

ABSTRACT

All contributors to *The Proceedings of the 2023 Aarhus International Conference on Voice Studies* examine the human voice from multiple disciplinary perspectives. As the editors of these proceedings know well from their own work, such interdisciplinarity comes with a wide range of challenges: of collaboration, of communication, of institutional and infrastructural accommodation. In this introduction, we wish to examine the nature of these challenges, and to provide some suggestions about how they might best be negotiated. First, however, we provide a brief overview of voice studies: The interdisciplinary study of the human voice.

Key words: voice, voice quality, voice studies, cross-disciplinary studies, interdisciplinary studies

1. Why interdisciplinary approaches to the voice?

If we had to summarise the contributions to these proceedings in a single proposition, it would be that voices can communicate so much more than just words. Vocal variation can communicate affect and emotions [1–4], such as sadness, happiness, apathy, fear, or anger. Whether we align with others in terms of our opinions and outlooks is also something that can be reflected in how we speak. For instance, researchers have found vocal variation that signals disagreement [5–7], and generally positive or negative alignment with a range of sociopolitical groups [8–10]. But different types of voices are also associated with different sex categories [11, 12], gender identities [12, 13], and sexual orientation [12, 14 15]. Different voices are perceived as variably attractive by different groups of listeners [16–18] and speakers can adopt these vocal characteristics in the course of an interaction as well [19–21]. This means that we can also use vocal variation to attract (or repel!) others. Much research has been conducted in the area of charismatic speech: Why do we perceive certain speakers as charismatic? Why are some speakers perceived as having better leadership skills? Unsurprisingly, then, research shows that perceived charisma, too, is associated with vocal variation [22–24]. Even moral judgement is cued by the properties of our voices, such that someone may be assumed to be a bad person based merely on how they speak [25, 26]. Thus, the areas where the voice plays an important role seem endless, and the list above fails to include other important aspects of human existence, such as ethnicity [12, 27, 28], age [29, 30], health [29, 31–33], art and aesthetics [34, 35], etc. At the same time, the studies referenced above only focus on human vocal variation, although vocal variation plays an important role in non-human communication as well [36, 37].

The human voice, then, sends many signals about its speaker. These signals are made by the complex biology of the human vocal tract; motivated by the aims and needs of human communicators; and mediated by technologies, such as radios or smartphones. In addition, all vocal communication goes on against the background of social norms and conventions that guide their interpretation. Considering the many areas of our lives in which the human voice plays an important role, it is not surprising that a range of disciplines have approached – and often collaborated on – this topic. As represented by the contents of the present proceedings alone, a list of such disciplines would need to include aesthetics, ageing studies, anthropology, biology, business studies, communication studies, culture studies, forensics, gender studies, history, linguistics, literature, media studies, medicine, musicology, psychology, sociology, speech engineering, and speech technology (limiting ourselves to the nature of the *human* voice). These disciplines, and many more, contribute to the interdisciplinary research field of voice studies.

To give a sense of just how differently voice might be investigated within voice studies, and to be able to discuss some of the challenges that surround interdisciplinarity in this introduction, we will quickly recount some contributions from a single research project involving most of the editors of these proceedings: "'It's Not What You Said, It's How You Said It': An Empirical Approach to the Voice as the Outward Expression of Inner Character." Many of the project's contributions focused on fiction: One examined how the vocal performances of video game characters can contribute to storytelling and the building

© 2024 Míša Hejná et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

of virtual worlds [38]. Another proposed that a sick and broken voice quality is often used to metaphorize the moral corruption of fictional characters [39]. A third examined the voices of good and evil characters performed in live-action roleplaying [40]. A fourth found that vocal variation has been used to construct the femme fatale character in American cinematography since the inception of sound film [7]. Finally, we conducted a quantitative study of accents and character roles (e.g., hero, villain) in Disney films [41]. Of contributions that did not revolve around fiction, one examined the role of *f*0 and vowel formant dispersion in cis-gender female-male flirting interactions [20]. Another examined Donald Trump's political campaigning and argued that his highly informal use of voice quality – including pitch, speech rate, and rhythm – served to distance him from the formal and self-censoring elites [42].

These publications represent collaborations across multiple fields, including sociophonetics, media studies, gender studies, political science, and psychology. They also employ very different methods, both qualitative and quantitative. They have therefore demanded much interdisciplinary awareness from, and translation between, the editors of these proceedings. In the remainder of this introduction, we will examine what we have identified as four of the main and most general challenges for such interdisciplinary work on voice.

2. Challenges of interdisciplinary approaches to voice studies we identified on our path

The particulars of some challenges of interdisciplinary approaches to voice studies will undoubtedly depend on the specific combination of the fields in question, while others are likely more universal across any attempts at interdisciplinary collaboration. Our team found that a great deal of time and energy was required simply to make sure that we understand one another. It could also be a challenge to identify suitable publication outlets for the interdisciplinary work produced by the project. This was linked to issues of standards across disciplines – what types of data and what types of analyses count as "good" and why [43]? Last but likely not least, when planning follow-up projects through grant applications, we encountered a challenge we had come to know from submissions of manuscripts to journals: Interdisciplinarity is good and wanted, but it would seem that this is the case only with some combinations of disciplines, and both journal and grant applications platforms are ultimately not generally designed to accommodate interdisciplinary work. We will elaborate on each of these challenges in what follows.

Communication between disciplines cannot be underestimated in terms of how time-consuming this can become. In our experience, even scholars who have previously collaborated still need to schedule more time into the planning process of a publication than if they collaborated only with scholars from their own subfields. This includes not only the planning stages of a project, but also data processing and analysis, the interpretation of the results, and the actual writing of the intended publication. On a fairly basic level, different disciplines come with different traditions of formulating research questions and/or arguments, and negotiating the structure of a paper is something that requires extra time. In addition, the analytical steps involved, and the time these require, may not be particularly clear to all team members involved, warranting more discussion. On a less basic level, different disciplines also come with different theoretical and analytical frameworks, and these may be potentially challenging to reconcile. This is linked not only to how different team members use specific terms but also to what is considered "good" data and "good" analytical procedures. Over a period of four years, we have regularly had to debate our understanding and usage of specific voice-related terms to ensure transparency across disciplines in a way which reconciles our need for academic precision and comprehension across disciplines. The way we tackled such issues was by investing a considerable amount of time into sending ideas, synopses, and drafts of our output back and forth between each other, i.e. not on producing novel analyses or more text of the article in question. Unfortunately, most academic institutions are not set up in a way that would value this "extra" time as work. Challenging communication across disciplines can take more practical forms as well. For instance, as someone working as a phonetician within the humanities, Míša has been attempting to set up an endoscopy system, which is predominantly found in medical areas. Two years after purchasing the system, Míša has not been able to find directions concerning the safe use and maintenance of the equipment. While these aspects may seem rather banal, not being based in a medical discipline has turned the endeavour into a task which is likely going to take years to resolve satisfactorily.

Although it has become popular to do interdisciplinary research, most journal outlets are either not particularly interdisciplinary, or they are interdisciplinary in rather specific ways – it is not uncommon for journals to have a primary disciplinary focus which may then be secondarily supplemented by the theory and methods of another discipline. Therefore, our strategy with journal publication, which we deem to have been successful, has been one where one discipline dominates and other disciplines are consulted for specific aspects of the publication. But even with all of this strategising, it was often quite difficult to tell which journals would be receptive. This resulted in the rejection of a few submissions, as well as a lot of careful examination of the fit between our output and the standards of different journals. This is again something that takes quite a lot of time, which needs to be acknowledged and considered.

But even with our strategy to turn various disciplines into ancillary ones when working on specific manuscripts, we repeatedly found ourselves facing other issues as well. A prominent issue has been a perceived tension between the quantitative and the qualitative. Phonetic sciences, and even sociolinguistics, have revealed quantitative biases. For instance, the 20th International Congress of Phonetic Sciences held in 2023 saw a special session on qualitative research in phonetics. Many phoneticians are well aware of the necessary inclusion of statistical analyses in phonetics papers. The organisers of the 50th conference on New Ways of Analysing Variation (a

sociolinguistic conference) have issued a statement on the website of the conference to assure authors that qualitative research was not penalised during the reviewing process, despite what some reviewers may have communicated in their evaluations of such abstracts (https://nwav50.stanford.edu/frequently-asked-questions/faq-abstract-review). The "It's Not What You Said, It's How You Said It" project was interesting in that it brought disciplines together in which the quantitative and the qualitative are valued differently. This very much affected our choices for journal outlets, which in turn meant that only partial integration of disciplines was practically possible.

Other than finding ourselves debating what would be perceived as good data by which journal, we also repeatedly discussed different disciplinary practices associated with different types of publications. Are proceedings contributions less prestigious than journal articles? That depends on the discipline. In some fields, such as computer science or engineering, publication in a conference proceedings can be just as highly regarded as publication in a journal article. By contrast, the proceedings format is generally not very widely used or highly regarded in humanities fields such as literary studies or media studies. Since different publication venues will therefore be incentivized for researchers of different disciplinary backgrounds, this can create difficulties when it comes to deciding where to submit a manuscript. With interdisciplinary teams working with different traditions of "prestige" associated with publication types, where do we decide to publish and why? Which choices are more strategic for the more junior members on the team to boost their options of securing a more stable job? What should the order of the co-authors be? In some disciplines, the place of the last author is seen as more prestigious, while in others it is the first or as close to the first as possible. Which order should one pick so that the right implications are secured across readerships identifying with diverse disciplines? Does one do whatever is best for the most junior members on the team? And how do different disciplines implement what is known as "peer-review"? These questions, we believe, should not impact our choices in where we decide to publish specific articles, but they did. This is the reality of interdisciplinary work.

Beyond the compromises we have made regarding the degree of interdisciplinary integration, various other challenges remain. In the many quantitatively biased journals for phonetic and sociolinguistic research, even a partial interdisciplinary integration means that more space is needed to make sure the readership can follow. But journals are strict with the maximum word counts allowed. In our experience, the more integrated a study is in terms of the diverse disciplines, the more space is needed to make the work as transparent to different types of readerships as possible. After all, the main point of interdisciplinary research *is* to communicate across disciplines.

Not only publication systems, but also grant application systems can cause problems for interdisciplinary research. When our team applied for various grants, we were often forced to choose a primary discipline in the grant submission system, even where we were allowed to indicate that the project was interdisciplinary in nature. Does this have consequences for the assessment of grant applications? Likely. As we have learnt, different disciplines have different standards of "good" data and "good" analyses, and different disciplines can be associated with different assumptions on a level as basic as how an argumentation structure is presented. So far, we have very much focused on interdisciplinary research, but similar issues are found in interdisciplinary approaches to teaching.

Are these challenges worth all the effort involved, especially in a world in which there does not seem to be enough time within the day to do all the tasks that we need to take care of? Despite the numerous challenges identified by our team above, we do think that this is all worth it, at least in some cases. We elaborate on this conviction in what follows.

3. Interdisciplinarity is well worth the effort

We started off our interdisciplinary voice project "It's Not What You Said, It's How You Said It" with the belief that interdisciplinarity is worth the trouble involved. Despite the challenges reported in this introduction, we have emerged at the end of the project believing this even more strongly than before. Firstly, it was the interdisciplinary effort involved in discussing the challenges sketched above that made us more aware of our own individual disciplinary biases. This has helped us to learn not only about other disciplines, but also those in which we had already entered the project as specialists. The interdisciplinary approach has given us a chance to step aside to inspect our own disciplinary homes and discover their strengths and weaknesses. But mostly, we simply believe that the work is necessary and worthwhile because our subject matter does not fit into neat disciplinary boxes — the world around us and human existence are, after all, fantastically complex. We therefore remain strong proponents of integrated work enabling us to arrive at new insights.

Thus, we see an important goal of the present proceedings as raising awareness of, and promoting collaboration within, the wide range of cross-disciplinary approaches to voice studies. This is the reason we founded the interdisciplinary Centre of Voice Studies at Aarhus University (https://cc.au.dk/en/centre-of-voice-studies), Denmark. The Centre hosts monthly online talks, now partially available via the Centre's YouTube channel (https://www.youtube.com/@CentreofVoiceStudies). We aim to expose the Centre's members and affiliates to a diverse range of approaches to voice studies and to engage these talks with an open mind. One of the aims of the talks is also to discuss matters relating to disciplinary differences, including not only methodology but also what types of questions are seen as worthwhile to ask and why within each field in question. The Centre has also hosted the 2023 Aarhus International Conference on Voice Studies. It was at this conference that many of the research contributions of the present proceedings were first presented.

The proceedings are divided into four areas:

© 2024 Míša Hejná et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

- Sociolinguistics
- Media and popular culture studies
- Linguistic typology
- Digital studies

Within these, we find variably disciplinary and interdisciplinary approaches to voice studies, with a range of diverse research questions and methodological approaches involved. Being a highly interdisciplinary area, voice studies faces a lack of consensus when it comes to something as seemingly basic as the terminology used to describe aspects of vocal variation, and the definition of voice is itself quite varied. In order to start building meaningful bridges between disciplines, we first need to identify what lies on each disciplinary island (or archipelagos of specific disciplines that have been integrated for a bit longer) to see which bridges would actually be useful. Because not many publication outlets are open to highly interdisciplinary approaches, another aim of the proceedings is to provide an outlet to interdisciplinarians working on voice studies who are eager to learn from one another.

4. Acknowledgments

We would like the thank the Carlsberg Foundation for generous research support (Young Researcher Fellowship, grant number 35891).

5. References

1. Pratt, T. (2023). Affect in sociolinguistic style. *Language in Society*, *52*(1), 1–26. https://doi.org/10.1017/S0047404521000774
2. Starr, R. L. (2015). Sweet voice: The role of voice quality in a Japanese feminine style. *Language in Society*, *44*(1), 1–34. https://doi.org/10.1017/S0047404514000724.
3. Starr, R. L., Wang, T., & Go, C. (2020). Sexuality vs. sensuality: The multimodal construction of affective stance in Chinese ASMR performances. *Journal of Sociolinguistics*, *24*(4), 492–513. https://doi.org/10.1111/josl.12410.
4. Yanushevskaya, I., Gobl, C., Ní Chasaide, A. (2018). Cross-language differences in how voice quality and f0 contours map to affect. *Journal of the Acoustical Society of America*, *144*(5), 2730–2750. https://doi.org/10.1121/1.5066448
5. Freeman, V. (2015). *The phonetics of stance-taking*. Dissertation, University of Washington. http://hdl.handle.net/1773/34001.
6. Grivičić, T., & Nilep, C. (2004). When phonation matters: The use and function of *yeah* and creaky voice. *Colorado Research in Linguistics*, *17*(1), 1–11. https://doi.org/10.25810/36ed-z920
7. Hejná, M., & Eaton, M. (in press). 'It's chiefly your eyes I think, and that throb you get in your voice': The place of creaky voice in the soundscape of attractive female voices in twentieth and twenty-first century American cinematography. In F. Venturi (Ed.), *Creak: Theories and practices of pulse phonation*. Jenny Stanford Publishing.
8. Jespersen, A. (2018). Innovations in the stylistic variation of nuclear tunes in Belfast English. *Proceedings of Speech Prosody 2018*, 527–531. https://doi.org/10.21437/SpeechProsody.2018-107
9. Hall-Lew, L., & Trousdale, G. (2020). Breksit or Bregzit: when political ideology drives language ideology. *University of Pennsylvania Working Papers in Linguistics*, *26*(2), 89–98. https://repository.upenn.edu/pwpl/vol26/iss2/11
10. Hall-Lew, L., Friskney, R., & Scobbie, J. M. (2017). Accommodation or political identity: Scottish members of the UK Parliament. *Language Variation & Change*, *29*(3), 341–363. https://doi.org/10.1017/S0954394517000175
11. Dauphinais Civitello, A. (2024). The liminal (vowel) space of womanhood: Fundamental frequency, formants, and the intersex body in Brazil. *Language in Society*. Advance online publication. https://doi.org/10.1017/S0047404524000319
12. Podesva, R., & Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, *35*, 173–194. https://doi.org/10.1017/S0267190514000270
13. Podesva, R. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, *11*(4), 478–504. https://doi.org/10.1111/j.1467-9841.2007.00334.x
14. Zimman, L. (2018). Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language & Linguistics Compass*, *12*(8), e12284. https://doi.org/10.1111/lnc3.12284
15. Sulpizio, S., Fasoli, F., Maass, A., Paladino, M. P., Vespignani, F., Eyssel, F., & Bentler, D. (2015). The sound of voice: Voice-based categorisation of speakers' sexual orientation within and across languages. *PLoS ONE*, *10*(7), e0128882. https://doi.org/10.1371/journal.pone.0128882
16. Boyd, Z., Fruehwald, J., & Hall-Lew, L. (2021). Crosslinguistic perceptions of /s/ among English, French, and German listeners. *Language Variation and Change*, *33*(2), 165–191. https://doi.org/10.1017/S0954394521000089
17. Coupland, N., & Bishop, H. (2007). Ideologised values for British accents. *Journal of Sociolinguistics*, *11*(1), 74–93. https://doi.org/10.1111/j.1467-9841.2007.00311.x
18. Šturm, P., Hejná, M., & Tylečková, L. (2024). Perceived vocal attractiveness of breathiness in Danish. Manuscript submitted for publication.
19. Anolli, L., & Ciceri, R. (2002). Analysis of the vocal profiles of male seduction: From exhibition to self-disclosure. *The Journal of General Psychology*, *129*(2), 149–169. https://doi.org/10.1080/00221300209603135
20. Habekost, E., Boyd, Z., & Hejná, M. (2024). The role of F0 and Vowel Formant Dispersion in cis-gender female-male flirting interactions. *34th Proceedings of Fonetik*.
21. Michalsky, J., & Schoormann, H. (2022). Birds of feather flock together but opposites attract! On the interaction of F0 entrainment, perceived attractiveness and conversational quality in dating conversations. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. Ohala (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers* (pp. 221–248). Springer. https://doi.org/10.1007/978-981-15-6627-1_12
22. Niebuhr, O. (in press). How foreign accents affect perceived speaker charisma. *SpeakOut*, *65*, 1–10.
23. Niebuhr, O., Brem, A., Michalsky, J., & Neitsch, J. (2020). What makes business speakers sound charismatic? A contrastive acoustic-melodic analysis of Steve Jobs and

Mark Zuckerberg. *Cadernos de Linguistica*, *1*(1), 1–40. https://doi.org/10.25189/2675-4916.2020.V1.N1.ID272
24. Siegert, I., & Niebuhr, O. (2021). Case report: Women, be aware that your vocal charisma can dwindle in remote meetings. *Frontiers in Communication*, *5*, 611555. https://doi.org/10.3389/fcomm.2020.611555
25. Dixon, J. A., Mahoney, B., & Cocks, R. (2002). Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology*, *21*(2), 162–168. https://doi.org/10.1177/02627X02021002004
26. Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United States*. Routledge.
27. Halliday, N. (2023). Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, *8*, 1116955. https://doi.org/10.3389/fcomm.2023.1116955
28. Moisik, S. R. (2013). Harsh voice quality and its association with blackness in popular American media. *Phonetica*, *69*(4), 193–215. https://doi.org/10.1159/000351059
29. Hwan Ryu, C., Han, S., Lee, M.-S., Yoon Kim, S., Yuhl Nam, S., Roh, J.-L., Ryu, J., Jung, J-S., & Choi, S.-H. (2015). Voice changes in elderly adults: Prevalence and the effect of social, behavioral, and health status on voice quality. *Journal of the American Geriatrics Society*, *63*(8), 1608–1614.
30. Spazzapan, E. A., Cardoso, V. M., Fabron, E. M. G., Berti, L. C., Brasolotto, A. G., & de Castro Marino, V. C. (2018). Acoustic characteristics of healthy voices of adults: From young to middle age. *CoDAS*, *30*(5), 1–7. https://doi.org/10.1590/2317-1782/20182017225
31. Hazlett, D. E., Duffy, O. M., & Moorhead, S. A. (2011). Review of the impact of voice training on the vocal quality of professional voice users: Implications for vocal health and recommendations for further research. *Journal of Voice*, *25*(2), 181–191. https://doi.org/10.1016/j.jvoice.2009.08.005
32. Luschei, E. S., Ramig, L. O., Baker, K. L., & Smith, M. E. (1999). Discharge characteristics of laryngeal single motor units during phonation in young and older adults and in persons with Parkinson disease. *Journal of Neurophysiology*, *81*(5), 2131–2139. https://doi.org/10.1152/jn.1999.81.5.2131
33. Rainho Rocha, B., & Behlau, M. (2018). The influence of sleep disorders on voice quality. *Journal of Voice*, *32*(6), 771.e1-13. https://doi.org/10.1016/j.jvoice.2017.08.009
34. Green, T. (2021). Through the Mercure's lens: Mid-eighteenth-century acting styles and vocal aesthetics at the Paris Opéra. In G. Sadler, S. Thompson, & J. Williams (Eds.), *The operas of Rameau: Genesis, staging, reception* (pp. 264–277). Routledge.
35. Neumark, N., Gibson, R., & van Leeuwen, T. (2023). *VØ1CE: Vocal aesthetics in digital arts and media*. MIT Press. https://doi.org/10.7551/mitpress/9780262013901.001.0001
36. Pisanski, K., Bryant, G.A., Cornec, C., Anikin, A., & Reby, D. (2022). Form follows function in human nonverbal vocalisations. *Ethology Ecology & Evolution*, 34(3), 303–321. https://doi.org/10.1080/03949370.2022.2026482
37. Puts, D. A., Hill, A. K., Bailey, D. H., Walker, R. S., Rendall, D., Wheatley, J. R., Welling, L. L. M., Dawood, K., Cárdenas, R., Burriss, R. P., Jablonski, N. G., Shriver, M. D., Weiss, D., Lameira, A. R., Apicella, C. L., Owren, M. J., Barelli, C., Glenn, M. E., & Ramos-Fernandez, G. (2016). Sexual selection on male vocal fundamental frequency in humans and other anthropoids. *Proceedings of Royal Society B*, *283*, 20152830. http://dx.doi.org/10.1098/rspb.2015.2830
38. Kjeldgaard-Christiansen, J., & Hejná, M. (2023). The voices of game worlds: A sociolinguistic analysis of *Disco Elysium*. *Games & Culture*, *18*(5), 578–597. https://doi.org/10.1177/15554120221115396
39. Kjeldgaard-Christiansen, J., Hejná, M., Clasen, M., & Eaton, M. (2023). Evil voices in popular fictions: The case of *The Exorcist*. *The Journal of Popular Culture*, 56(2), 226–247. https://doi.org/10.1111/jpcu.13234
40. Boyd, Z., & Hejná, M. (in press). The 'Critical Role' of acoustic and auditory methods in holistic studies of voice quality: A case study of non-player characters in Dungeons & Dragons voiced by Matthew Mercer. *Language in Society*.
41. Kjeldgaard-Chrstiansen, J., Boyd, Z., & Hejná, M. (in press). Teaching children to discriminate? A quantitative study of linguistic representation in Disney's 'Revival Era' animated films. *MedieKultur: Journal of Media and Communication Research*.
42. Kjeldgaard-Christiansen, J. (2024a). The voice of the people: Populism and Donald Trump's use of informal voice. *Society*. Advance online publication. https://doi.org/10.1007/s12115-024-00969-7
43. Kjeldgaard-Christiansen, J. (2024b). What science can't know: On scientific objectivity and the human subject. *Poetics Today*, *45*(1), 1–16. https://doi.org/10.1215/03335372-10938579

© 2024 Míša Hejná et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-002

Intersections of sexual orientation, bilingualism, and /s/ variation in French and German men's speech

Zac Boyd[1]

[1]Department of English, Aarhus University, Denmark

contact@zacboyd.co.uk

ABSTRACT

Previous studies of the interaction between language, gender, and sexual orientation have shown /s/ to be a robust variable indexing gay identity in men's speech. This study expands upon the previously established linguistic framework of indexing gayness by exploring /s/ variation in native and non-native speech of bilingual speakers, examining how the linguistic construction of gay identity interacts between L2 English production and constraints of their native language based on read speech data of 19 gay and straight French and German men. Results show that some gay speakers produce /s/ with a higher centre of gravity (CoG) than the straight speakers. Results also suggest that, overall, a speaker's native language does not have a significant impact on the quality of his /s/ production in English. These results are consistent with previous findings, which show sibilant variation to index sexual orientation in monolingual gay men's speech and provide evidence of this feature acting as an index of sexual orientation in French and German. Furthermore, the results call for a greater level of inquiry into how gay speakers who employ this feature construct their gay identities beyond a purely gay/straight dichotomy.

Key words: sexual orientation, /s/ variation, bilingualism

1. Introduction

This article explores sibilant variation between native language (L1) and second language (L2) speech examining the effects of sexual orientation in bilingual English speaking French and German men at the intersection of second language acquisition (SLA) and sociolinguistics. This intersection has been a notable trend in SLA research, particularly examining sociolinguistic competence [77] and applications of first wave [41-43] and third wave [26-27] approaches to sociolinguistics to gain insight into social processes of a speaker's interlanguage [46]. Following previous variationist SLA studies, this paper examines the potential for cross-linguistic social meaning through phonological variation of gay bilingual men's speech.

The indexical associations of gay men's speech have been widely studied in sociolinguistic research, establishing the generally accepted notion that there is some quality of certain gay men's speech that signals sexual orientation (e.g., [44]; [53]; [57]). Much of this work aims to discover which phonetic features signal gayness in both speech production and speech perception in monolingual speakers ([13]; [15]; [33]; [40]; [58]; [60-61]; [69]). However, research in this area regarding bilingual speakers is still quite lacking (though see [30]; [35]; [36]; [83]). This study addresses this gap by examining one key variable that indexes sexual orientation and gay identity in men: the voiceless sibilant /s/.

This paper analyses read speech data from both the native language (L1) and second language (L2) of nineteen gay and straight French and German men, with L2 referring to English spoken by these individuals. The findings indicate that some gay French and German speakers exhibit /s/ variation similar to that observed in monolingual gay men's speech, characterized by a higher centre of gravity (CoG) /s/ production compared to straight speakers, both in their L1 French or German and L2 English productions. This variation is speaker-dependent and occurs cross-linguistically for both French and German gay men, transcending nationality and language boundaries. These findings challenge a simple gay/straight dichotomy and call for a more nuanced understanding of the diverse constructions of gay identities available to speakers. The current results call for further exploration of the complicated interplay of the many factors that culminate in the production of these multilingual, cross-cultural indices. In this, I argue that looking at the gay speakers of the present work as a single homogeneous group does not fully explain the linguistic variation seen here, and calls for a more in depth understanding of the construction of the multiplicity of gay identities available to speakers.

© 2024 Zac Boyd. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Theoretical framework

2.1. Gay speech style and /s/ variation

This study examines stylistic variation of /s/ in relation to sexual orientation in non-native English speakers. Sibilant variation is a prominent feature of a gay speech style and is the primary focus here. Campbell-Kibler et al. [17] emphasize that styles convey social meaning, and while the full definition of a gay speech style remains uncertain [44, pp. 533], prior research has consistently linked /s/ variation to gay identity and non-normative masculinity in English ([16]; [23]; [44-45]; [48]; [52]; [53]; [54]; [60-61]; [62]; [65]; [70-71]; among others). While /s/ is not the sole marker of a gay speech style, it is one of the most recognizable features, and is seen across various languages (e.g., Afrikaans [5], Danish ([49]; [56]; [57]), Hungarian [63], and Spanish ([30]; [47]; [78]).

Nevertheless, higher frequency /s/ production is not exclusive to a gay speech style. For instance, (largely cisgender white) women exhibit significantly higher /s/ frequencies than men, driven by both biological and social factors ([31]; [32]; [38]; [73]). Additionally, hyperarticulation of /s/ correlates with perceived speech clarity, level of education, and formality [53].

Beyond relatively few studies on French ([35]; [36]; [67]) and German ([34]; [39]), there is little work examining the phonetic correlates of a gay speech style in these languages, despite evidence suggesting that spectral aspects of /s/ variability convey similar social meanings across languages. For instance, Mack [47] finds in Puerto Rican Spanish and Pharao et al. [57] in Danish, that speakers exhibiting /s/ variation with similar social associations as English. Rácz and Schepácz [63] link higher sibilant frequencies to "feminine-sounding speech" in male speakers of Hungarian. Russell [67] observes higher /s/ centre of gravity (CoG) in overtly performative French speech associated with sounding gay, while Kachel, Simpson, and Steffens [39] report a similar trend in German gay speakers producing slightly higher /s/ CoG.

These studies largely involve monolingual speakers/listeners, with limited exploration of sociolinguistic indexicality and second language acquisition in bilingual queer speakers. Fisher's [30] case study of a gay bilingual Spanish/American English speaker shows stylistic variation of /s/ between L1 Spanish and L2 English indexing sexual orientation both in L1 Spanish and (for this speaker) potentially even more so in his L2 English via longer /s/ duration and relative high CoG frequency. Much like the present paper, Hobart [35-36] looked at bilingual gay and straight French men. Only in one of these papers [35] does he show that gay French speakers produce /s/ with a higher CoG and more negative skewness than the straight speakers. This was seen for both the L1 and L2 to varying degrees. However, in his later paper CoG and skewness were not predicted by sexual orientation at either level [36].

The gay speech style discussed here is socially conditioned, meaning that it is not a style used by all gay men ([14]; [54, pp. 203]), nor are features associated with it used exclusively by gay men (i.e., /s/ variation). A style may possess a wide number of indexically meaningful features embedded within the speech signal but features are not bound to a single indexical meaning, rather any linguistic feature may have an unlimited number of potential meanings dependent on contextual renegotiation, something Eckert [27] refers to as 'indexical mutability'. It is through this process which a gay speech style may emerge. As such, /s/ variation is only one of a wide number of potential linguistic factors which may aide in indexing a gay speech style within the speech of any individual.

2.2. Bilingualism and /s/ for French/German L2 English speakers

This paper focuses on individuals who would be classified as late bilinguals, where they had not fully acquired their second language (L2), English, prior to the end of the Critical Period of language acquisition. The critical period ends around adolescence ([25]; [37]) and suggests that younger learners have an advantage in acquiring native-like pronunciation. The participants of the present study started learning English between ages 8-22, well past the critical period, and did not achieve fluency until later in their education (if at all). While it has been suggested that late bilinguals may need to exert a more conscious effort to produce L2 phonology due to native language (L1) influence [80, pp. 220], the sibilant systems in French, German, and English are minimally different. Therefore, participants should not encounter issues of single category assimilation in their sibilant productions. While a full detailed discussion of the sibilant structures of these languages is beyond the scope of this paper, the similarities between the three allow us to explore how /s/ may signal shared social meaning across languages.

3. Methods

The data used in the analysis comes from individual recording sessions with each participant. All recordings took place in 2015, resulting in a total of 19 speakers between the ages of 20 and 30 (mean age 26.7). While not explicitly controlled for, all speakers fall into a relatively narrow demographic of white, middle class, highly educated (undergraduate or higher), cisgender men. Speakers were recorded at 24-bit quantization with 48 kHz sampling frequency in public (but quiet) locations in each city (Paris, Lyon, Berlin, and Düsseldorf).

Data for the present study comes from read speech from both the speaker L1 and L2. To elicit speech from the participants' native language, participants read a native language version of Little Red Riding Hood: *Le Petit Chaperon Rouge* for French and *Rotkäppchen* for German. For English speech, participants were asked to read part of an English language translation of Snow-White. Speakers averaged approximately eight minutes per reading task. These passages were selected to ensure that each passage was of comparable length and literary style between all languages examined in this paper.

All data underwent a high-pass filter at 1,000Hz which helps to remove unwanted noise within the recordings and aides in maintaining a frequency threshold of approximately 4,000-10,000Hz, a range at which most /s/ variation occurs [31]. Each file was transcribed, underwent forced alignment using the FAVE program suite [66], and was extracted in Praat [7]. Prior

to analysis, outlier tokens measuring ±2 standard deviations (separated per speaker), any tokens measuring less than 30ms, tokens with obvious Praat tracking errors, and all /STR/ clusters (known to retract; see [3]) were removed. This resulted in an average token count of 347 for each of the French speakers (L1 n=174; L2 n=173) and 425 for each of the German speakers (L1 n=251; L2 n=174).

3.1. Perceived nativeness ratings

To assess the impact of English language proficiency on L2 speech production, a modified methodology based on Sorace and Filliaci [72] was used. English language ability was rated by thirty-one L1 English speaking respondents. Speakers underwent evaluation on six factors: syntax, morphology, lexicon, phonology, fluency, and overall impressions. This method of assessing English language ability should be viewed as a proxy of proficiency as respondents were not directly asked to rate a speaker's proficiency, but rather their perceived level of 'native-like speech'. Overall, the French speakers were rated as less proficient than German speakers with average total ratings of 61.78/100 and 70.82/100 respectively.

The effect of English proficiency is established by comparing two linear mixed effects models in a likelihood ratio test. Table 1 shows the comparison of two models: 1) CoG ~ Orientation + Language + Orientation:Language + (1|Speaker) and 2) this same model with the inclusion of 'Phonology' from the proficiency evaluations. In all models, 'Language' refers to either L1 French, French speaking L2 English, L1 German, or German speaking L2 English.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) show differences between the two models with the smaller number representing a better model fit. The AIC shows no difference between the two models and the BIC slightly favours the model which does *not* include the English proficiency rating for 'Phonology'. The likelihood ratio test therefore indicates that phonological proficiency does not improve the model fit.

Table 1: *comparison of models examining the effect of phonological proficiency.*

	df	AIC	BIC	logLik	deviance	χ2	χ2 df	p
Orientation* Language	10	264713	264790	-132346	264693			
+Phonology	11	264713	264798	-132346	264691	1.3608	1	0.2434

The results of the likelihood-ratio tests are consistent throughout all models for each different proficiency factor (Syntax, Morphology, Lexicon, etc...) and phonetic measure (in this case CoG, but also for skewness and Peak Frequency – not discussed within). Since the model fit (slightly) improves when English proficiency is excluded (as seen via the BIC), we must assume that English proficiency has no effect on /s/ variability, something which can likely be attributed to the minimal differences between the sibilant structures of these languages. As such, any differences which arise between L1 and L2 speech can likely be attributed to aspects of the speech elicitation methods, as opposed to inherent language differences or English proficiency (see discussion below and Boyd 2018b). Given that many speakers with very similar proficiency ratings have drastically different results in their /s/ productions, it is highly unlikely for proficiency to be qualitatively meaningful at the individual level for this specific variable and as such, the effects of English proficiency will not be discussed in further detail (for more information see [10, pp 29-34]).

4. Results

The data presented here is based on /s/ production data from L1 and L2 read speech for all 19 participants (token count *n*=7499). The analysis for the present paper is based on CoG, a measure of the weighted average frequency of the sibilant [82]. CoG is one of the primary measures considered in previous sociophonetic work on /s/ variation, specifically work on /s/ and sexual orientation. In the interest of space, results from the analysis of both Skewness and Spectral Peak Frequency are not presented here. However, in both instances, results are similar to what is seen in CoG (with skewness being inversely correlated with CoG) and Spectral Peak Frequency providing no further insights beyond what is seen in the CoG analysis (see Boyd [10] for a full discussion).

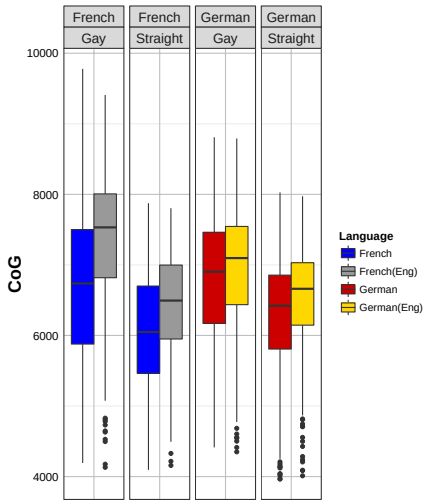


Figure 1: */s/ CoG by Orientation & Nationality.*

Figure 1 shows the distribution of /s/ CoG production data (represented in non-normalised Hz values) by nationality and sexual orientation. These plots show overall trends for higher CoG by gay men, and a trend across nationality and orientation for higher English productions. Furthermore, little overall differences are seen between L1 or L2 across nationality, with L1 French and German being similar, and both L2 Englishes having comparable production values, though the difference is slightly greater for the gay speakers. The French gay speakers show an overall average CoG Frequency of 6,670Hz in their L1 and 7,310Hz in their L2. This is compared to the straight French speakers who have an average CoG of 6,045Hz in the L1, and 6,426Hz in the L2. Comparatively the gay German speakers have an average CoG of 6,805Hz in their L1 and 6,965Hz in their L2,

© 2024 Zac Boyd. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

where the straight speakers show an average CoG at 6,422Hz for L1 speech and 6,615Hz for L2 speech.

Linear mixed-effects models were fit running the lme4 package [4] and confidence intervals were estimated via a bootstrapping process in the boot package ([18]; [24]). 95% confidence intervals were estimated from 5,000 parametric bootstrap replicates [4]. While bootstrapping models do not give a p-value they produce an output which replaces significance values in favour of a different estimate of the reliability of the effect. This output can be interpreted such that if the results of the confidence interval excludes zero the effect is considered to be a reliable predictor of this variation (or 'significant').

For the models, there are two variables that comprise three fixed effects: sexual orientation, language used (L1 French or German; French/German L2 English), and an interaction effect between the two. There are also two variables as random intercepts: speaker and target word. The reference level for each model is Orientation = Gay, Language = L1.

Figures 2 and 3 shows the results of the parametric bootstrap estimates and 95% confidence intervals for the mixed-effects model 'Centre Of Gravity ~ Orientation + Language + Orientation:Language + (1|Speaker) + (1|TargetWord)'.[1] For each of these figures "Sexuality Effect; *Nationality*" shows the difference between gay and straight speakers when speaking their L1. Similarly "Sexuality Effect; *Nationality*(Eng)" shows the differences between gay and straight speakers when speaking their L2. Finally, "Language Effect; *Sexual Orientation*" shows the differences between the L1 and L2 for gay or straight speakers.

These models indicate that for CoG there is a main effect of sexual orientation only for French speakers ('Sexuality Effect; French' and 'Sexuality Effect; French(Eng)'). The bootstrapping results show that straight speakers have a significantly lower /s/ CoG than the gay speakers, with /s/ productions of 652Hz lower when speaking in their L1. This difference is even greater when speaking English with the straight French speakers averaging 893Hz lower than the gay speakers. For German speakers there is no significant difference for a main effect of sexual orientation, though the model shows straight speakers have a slight overall trend for lower CoG than the gay speakers ('Sexuality Effect; German' and 'Sexuality Effect; German(Eng)'). Straight German speakers produce /s/ CoG 424Hz less than the gay German speakers in L1 German and 400Hz less in L2 English.

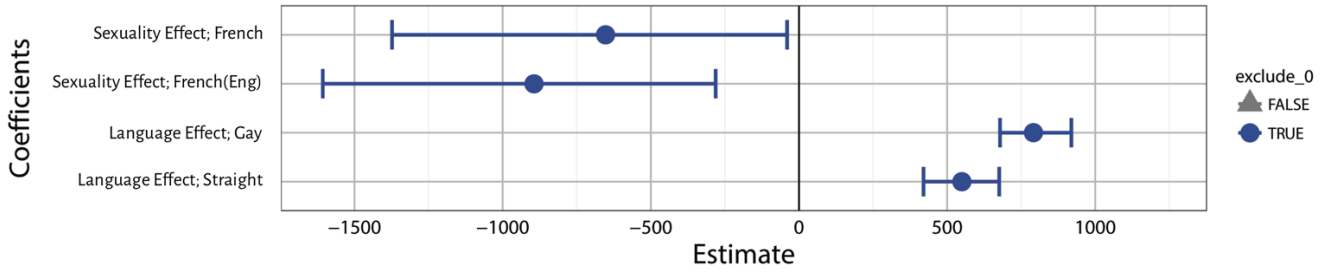


Figure 2: */s/ CoG: 95% bootstrapping Confidence Intervals - French speakers.*

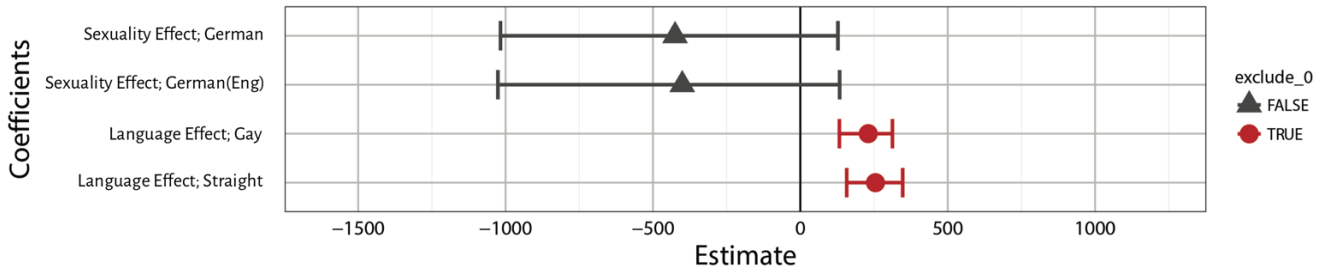


Figure 3: */s/ CoG: 95% bootstrapping Confidence Intervals - German speakers.*

The model also shows a reliable difference between these speakers' L1 and their productions of L2 English for both the French and German speakers ('Language Effect; (*Orientation*)'). For French speakers this difference is slightly larger than the German speakers. L2 English /s/ is produced with a higher CoG than in L1 French with straight speakers producing L2 English 549Hz higher on average. The gay speakers have an even greater difference between their L1 and L2 /s/ productions with English

[1] For exact confidence interval values for all results see Appendix 1.

791Hz higher on average than L1 French. German speakers show this same difference with L2 English /s/ being higher than L1 German by an average of 230Hz for gay speakers and 255Hz for straight speakers.

Mixed-effects models were also used to examine differences between the two nationalities. No significant results are seen in either the L1, nor the L2, regardless of sexual orientation when examining how these speakers differ from their national counterparts (i.e., gay L1 French compared to gay L1 German) indicating that /s/ production in L1 French and L1 German have little difference between them which is similarly seen in L2 English across both nationalities. Furthermore, no results are seen when accounting for regional variation within France or Germany.

These results have shown that there are strong similarities across nationalities as seen within both the descriptive statistics and the results of the mixed-effects models. The models outlined in this section show no differences between the two nationalities on either the L1 or L2. They also indicate that though gay French speakers may employ this higher /s/ CoG, the same cannot be said for the gay German speakers, though they do approach statistical reliability (according to the current models; e.g., Figure 3). One interpretation may be that sexual orientation really does not matter in relation to German /s/ productions, but it is also possible that there may be an effect which cannot be accurately captured within these models (see Boyd [9] for further discussion).

5. Discussion

This study set out to determine the potential for /s/ variation to correlate with sexual orientation within the speech of French and German men and if so, to investigate how this feature would manifest within second language speech. The results here present multiple findings of interest. First, it appears that this feature does indeed correlate with sexual orientation for French speakers in both their L1 as well as their L2, though, based on the inferential statistics, the same cannot be said of German speakers. Second, results have shown that the French and German men of this study, consistently produce /s/ differently between their L1 and their L2. Each of these findings have multiple explanations and interpretations which this section aims to address.

One argument that can be made here is that /s/ is being employed by some gay French and German men as a marker of their sexual orientation. This argument, however, is restricted to /s/ *production* among only a subset of the gay speakers and is not necessarily applicable to *perception* (as seen in [11]). Figure 4 succinctly delineates this point. This figure shows individual speaker mean values for CoG with each dot representing one individual speaker. The line between two dots shows the difference between the respective L1 and L2 production averages. A dotted line has been added to the figure by hand to illustrate the highest mean CoG of the straight speakers. While many gay speakers do not produce /s/ differently than the straight speakers, it is only the gay speakers who produce /s/ with a relative high CoG. In short, it is only a small number of the gay speakers who are doing anything of interest with their /s/ productions.

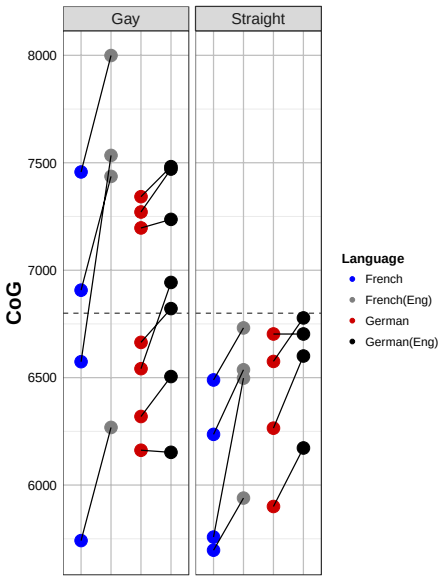


Figure 4: *speaker mean values - /s/ CoG.*

Why this specific finding arises can be understood by looking through the lens of hegemonic masculinity. Connell and Messerschmidt [22] suggest that masculinity, specifically hegemonic masculinity, is enacted heteronormative behaviour related to a male-female dichotomy. As such, any deviations within these two categories are delegitimised and relegated as subordinate ([20]; [74]).

In this, some gay men attempt to align with these constructs (e.g., 'straight-acting'), while others may (at least in part) ignore them (cf. [21]; [29]; [68]). When wholly discussing group level differences such as "gay men do 'this' and straight men do 'this'", we inherently assume group level homogeneity. However, this assumption is problematic and any assumption of group homogeneity, just because these speakers identify as gay, does not give a complete picture of the potential stylistic practices available to any one individual.

While Kachel, Simpson, and Steffens [39] show that gay German men tend to have higher /s/ CoG than the straight men of the study (a result which only approaches significance), the wider results of that study suggest greater differences within gay men and straight men than between them. The results provided here may then point to this same notion. Indeed, a large amount of diversity within wider groups should not be surprising. However, and given that there is very little research regarding French and German and /s/ variation in the context of sexual orientation, this research began by assuming a gay/straight divide to get at broader, surface level variation, aiming to determine if /s/ variability even correlates with sexual orientation as opposed to taking a position of in-group diversity at the outset. With such correlations seen within the current dataset it then becomes important to examine the motivations for this variation at the individual level (as discussed in great detail in [9]).

© 2024 Zac Boyd. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Figure 4 also highlights the second finding of interest, something also seen within the inferential statistics: /s/ variation between L1 and L2 speech. Due to the clear differences seen between the L1 and L2 for both nationalities it must be assumed that this is not a result of phonetic L1 language transfer [8]. I argue instead that these findings are a result of style shifting according to task type as seen via the *attention-paid-to-speech* [42-43] model of style shifting. Previous studies have shown that L2 reading is related to both language skills in the L2 and reading ability within the L1 (cf. [1]; [19]). An assumption of the latter would be pure conjecture on behalf of the researcher, but as previously mentioned, there were no relevant findings in terms of L2 proficiency which suggests that there is some underlying mechanism which motivates this shift beyond language used.

An explanation can be inferred from research on read speech and the gay speech style. Variation of sibilants, like all variation, can occur for a variety of different reasons. In a study of English fricatives, Maniwa et al. [50] compared acoustic measures between conversational and read speech. Results from that study show that across all speakers and sibilants (/s/, /ʃ/, and /z/) the clear-speech style contained intentionally hyperarticulate sibilants with longer duration, higher centre of gravity, higher spectral peak frequencies, and lower skew [50, pp. 3968], with similar results found in word list speech by Munson, et al. [54]. Furthermore, Munson [52] critiques Van Borsel et al. [76] in their use of read speech as the only basis for analysis because the /s/ variation shown by gay men in that study "may have had no relationship to social identity, but may have simply been due to group differences in the use of clear-speech variants" [52, pp. 4]. The L1 and L2 reading passages used in the present study could be predicted to act in a similar fashion, with the L2 passage resulting in variants akin to the hyperarticulate 'clear-speech variants'. L2 reading, especially if speakers are not fluent, requires more effort, or at least a different set of skills on behalf of the speaker [79]. Labov has suggested that styles may emerge based on attention paid to task type [43]. If one interprets L2 reading as a more 'formal' or complex task than L1 read speech, the results presented here of L2 English read speech being produced with higher CoG than L1 read speech are supported by previous studies suggesting differences on a spectrum from less to more formal speech tasks.

Furthermore, Brantmeier states that an 'L2 reader brings many characteristics (e.g., gender, background knowledge, experiences, interests, and personality) to the text' [12, pp. 34]. In other words, aspects of the reader's identity. Though it should be noted that her study is centred on reading comprehension, enjoyment, and recall as opposed to a sociolinguistic based approach of speaker identity. I agree with this view of read speech and argue that identity is not a switch which is turned on or off based upon the speech act, but rather the types of identity categories conveyed via read speech is just part of this other speaking style to be drawn on to convey a variety of social meanings. In essence, a read speech style may not only convey aspects of formality and speech clarity as discussed above, but may also convey salient identity categories (e.g., gender, social class, or sexual orientation), an argument supported by Podesva who suggests that the 'clear speech' style (distinctive of read speech) may in itself convey a variety of social meanings (e.g. 'competent', or 'prissy') depending on social context [59-60].

5.1. Cross-linguistic differences, bilingualism, and transnationalism

Though aspects of hegemonic masculinity and theories of style shifting (i.e. attention paid-to-speech models) help to explain the variation described above, this discussion would be incomplete without a greater acknowledgement of these speaker's bilingual identity and their relationship to the global gay community. An obvious starting point for this discussion entails sociolinguistic competence among late bilinguals which may also include the understanding and appropriateness of employing phonetic features which carry social meaning in the L2. Given that, for most English speakers, /s/ variation has a very strong meta-discursive function in indexing gayness (c.f., the documentary *Do I Sound Gay?* [75]), it may be the case that the gay speakers with significantly higher /s/ CoG productions have gained such a level of sociolinguistic competence to employ fronted /s/ variants to index a gay identity in L2 English. Though further work [9] suggests the speakers of this study may be style shifting in English (one marker of sociolinguistic competence [64]), it does not appear that these speakers have actually attained sociolinguistic competence. These speaker's report very little interaction with native English speakers in their day-to-day lives, they reported using English in extremely limited contexts (such as giving tourists directions or consumption of American media), they have never lived in a country where English is the *lingua franca*, and they have no explicit social motivation to gain native-speaker-like patterns of variation. In short, being French or German in France or Germany does not require socio-indexical knowledge of English. Furthermore, were some gay speakers to have acquired the sociolinguistic competence to employ fronted /s/ variants in their L2 as an index of gayness, the implication would then be that they have transferred that back to their L1, something which remains very unlikely.

Variationist sociolinguistics often focuses on speech patterns in a localised context. However, the effects of globalisation may transform a community such that the lines between the 'local' and the 'global' are blurred. Papastergiadis argues that "people now feel they belong to various communities despite the fact that they do not share a common territory with all other members" [55, pp. 115]. Based on the present data, it is impossible to determine if this fronted /s/ variant arose independently in France and Germany, if it is a product of cross-linguistic (transnational) cultural transfer from a language such as English, or an amalgamation of the two. Eckert [28] argues that to situate social meaning in a community we must look at not only the differences between groups within a given community nor differences between two geographically dissimilar communities, but rather both of these in context. An apt term in helping to explain this cross-linguistic sibilant variation comes from Blackwood referring to "Cultural Location", which establishes "the global, regional, and historical flows that have created specific discourses, knowledges, and ways of understanding the world in specific areas" [6, pp. 222] drawing on Appadurai's [2] idea of ethno-scapes that signify "the disjunctive flow of meanings" [6, pp. 222]. In the case of gay identities, the location becomes defined by a speaker identifying with these cross-

cultural cross-linguistic aspects of a global gay community within the confines of their geographical location, however the ideological representations of what that means to them become merged in local, national, and global aspects. This "disjunctive flow of meanings" could potentially be applied to the cross-cultural and cross-linguistic indexicality of certain features, namely sibilant variation.

A prime example of this comes from Zhang's study of 'Beijing Yuppies' which shows that speakers may draw on not only local linguistic resources but also extra-local and global sources to construct a new style [81]. Zhang's study shows that speakers can develop a "hybrid linguistic competence", and she argues that the language or language variety which a feature comes from is less important than "the *selective combination* of features from several varieties" [81, pp. 458-59, original emphasis] which are employed to give it new meaning. This may be similar to the speakers of the present study who report that their notions of gay stereotypes often come from archetypical gay characters portrayed in American media. If it is the case that these speakers are drawing on representations of gay identity in American media they may be constructing an identity via a process of bricolage (c.f. [26]). By combining semiotic resources of a gay style in France and Germany with those adopted from American media, speakers may be constructing a gay identity with no true origin.

6. Conclusion

The present study examined the role of /s/ variation and sexual orientation in read speech of French and German gay and straight men and has shown that some gay men may employ higher frequency /s/ variants which correlates with some aspect of their gay identity. This research has shown that if a speaker produced this feature (more fronted, higher frequency /s/ variants) in their L1 this is also seen within their L2 speech, supporting findings seen in the bilingual speaker in Fisher [30]. While a conclusive discussion of phonetic features which may index sexual orientation in French and German men is outside the scope of this discussion, this paper has determined /s/ variation to correlate with non-normative sexual orientation in French and German male speech. Furthermore, it has suggested several potential explanations for variation between the L1 and L2, and has explored various potential motivations for the production patterns seen here.

This research is part of a wider area at the cross-section of sociolinguistics and second language acquisition: determining if individuals employ /s/ variation in a way which correlates to sexual orientation in L1 and L2 speech production. Previous research in this specific area, gay identity in L2 speakers, is minimal. By focussing on non-native speakers this research fills a much needed gap in the literature on gay voices as the majority of previous research is based on monolingual speakers. I argue that many of the gay speakers of this study are exhibiting linguistic variation which is part of a distinct gay speech style. This feature, /s/ variation, is utilised across cultural and linguistic boundaries regardless of nationality or whether a speaker is using their L1 or L2. While this approach to the variation exhibited by these speakers may admittedly be an oversimplification of the nuanced distinctions that likely exist between (and within) the French and German speakers, it provides a grounded starting point for future study and has established that this may be a socially conditioned marked feature in French and German.

7. Acknowledgments

Thank you to Lauren Hall-Lew, Josef Fruehwald, Erez Levon, and Claire Cowie for their feedback on early versions of this work. Carlsberg Foundation Young Researchers Fellowship.

8. References

1. Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem. In J. C. Anderson & A. H. Urquhart (Eds.), *Reading in a foreign language*. New York: Longman.
2. Appadurai, A. (1996). *Modernity at large: Cultural dimensions of globalization*. Minneapolis: University of Minnesota Press.
3. Baker, A., Archangeli, D., & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change, 23*(3), 347–374.
4. Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting liner mixed-effects models using *lmer4*. *Journal of Statistical Software, 67*(1), 1–48.
5. Bekker, I., & Levon, E. (2017). The embedded indexical value of /s/-fronting in Afrikaans and South African English. *Linguistics, 55*(5), 1109–1139.
6. Blackwood, E. (2005). Transnational sexualities in one place: Indonesian readings. *Gen- der and Society, 19*(2), 221–242.
7. Boersma, P. & Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.17, retrieved 10 September 2023 from http://www.praat.org/
8. Bohn, O.S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 279–304.
9. Boyd, Z. (2018a). A certain kind of gay identity: [s+] and contextually mediated variation in bilingual French and German Men. Unpublished Manuscript.
10. Boyd, Z. (2018b). Cross-linguistic variation of /s/ as an index of non-normative sexual orientation and masculinity in French and German men. Doctoral Dissertation. The University of Edinburgh.
11. Boyd, Z., Fruehwald, J., & Hall-Lew, L. (2021). Cross-linguistic perception of /s/ among French, German, and English listeners. *Language Variation and Change*, *33*(2), 165-191.
12. Brantmeier, C. (2003). Beyond linguistic knowledge: Individual differences in second language reading. *Foreign Language Annals, 36*(1), 33–43.
13. Cameron, D. (1997). Performing gender identity: Young men's talk and the construction of heterosexual masculinity. In S. Johnson & U. H. Meinhof (Eds.), *Language and Masculinity* (pp. 47–64). Oxford: Blackwell.
14. Cameron, D., & Kulick, D. (2003). *Language and sexuality*. Cambridge University Press.

© 2024 Zac Boyd. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

15. Camp, M. (2009). *Japanese lesbian speech: Sexuality, gender identity, and language*. Unpublished Doctoral Dissertation. The University of Arizona.
16. Campbell-Kibler, K. (2011). Intersecting Variables and Perceived Sexual Orientation in Men. *American Speech, 86*(1), 52–68.
17. Campbell-Kibler, K., Eckert, P., Mendoza-Denton, N., & Moore, E. (2006). The elements of style. Poster Presented at *New Ways of Analysing Variation 35*.
18. Canty, A., & Ripley, B. (2017). boot: Bootstrap R (S-Plus) Functions. R Package version 1.3-20.
19. Chuang, H. K., Joshi, R. M., & Dixon, L. Q. (2012). Cross-Language Transfer of Reading Ability: Evidence From Taiwanese Ninth-Grade Adolescents. *Journal of Literacy Research, 44*(1), 97–119.
20. Connell, R. W. (1987). Gender and power: Society, the person and sexual politics. Cambridge: Polity Press.
21. Connell, R. W. (1995). *Masculinities*. University of California Press.
22. Connell, R. W., & Messerschmidt, J. W. (2005). Hegemonic Masculinity: Rethinking the Concept. *Gender and Society, 19*(6), 829–859.
23. Crist, S. (1997). Duration of onset consonants in gay male stereotyped speech. *University of Pennsylvania Working Papers in Linguistics, 4*(3), 4.
24. Davidson, A. C., & Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.
25. DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–533.
26. Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics, 12*(4), 453–476.
27. Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology, 41*, 87–100.
28. Eckert, P. (2018). Meaning and Linguistic Variation: The Third Wave in Sociolinguistics. New York, NY: Cambridge University Press.
29. Eguchi, S. (2009). Negotiating Hegemonic Masculinity: The Rhetorical Strategy of "Straight-Acting" among Gay Men. *Journal of Intercultural Communication, 38*(3), 193–209.
30. Fisher, I. W. (2016). Transfer of stylistic phonetic variables indexing sexuality in second language contexts. Unpublished Masters Thesis. Kansas State University.
31. Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeny, J. (1999). Acoustic Characteristics of /s/ in Adolescents. *Journal of Speech, Language, and Hearing Research, 42*(3), 663–677.
32. Fuchs, S., & Toda, M. (2010). Do differences in male versus female /s/ reflect biological or sociophonetic factors? In S. Fuchs, M. Toda, & M. Zygis (Eds.), *Turbulent sounds: An interdisciplinary guide* (Vol. 21). Walter de Gruyter.
33. Gaudio, R. P. (1994). Sounding gay: Pitch properties in the speech of gay and straight men. *American Speech, 69*(1), 30–57.
34. Guzik, K. M. (2006). Acoustic analysis of phonetic parameters of less masculine sounding German speech. *Arbeitsberichte Des Instituts Für Phonetik Und Digitale Sprachverarbeitung* Universität Kiel, 36, 15–29.
35. Hobart, M. (2013). Manifestation of sexual identity in the speech of French second language speakers of English: Evidence from the gay /s/. Unpublished Master's Thesis. The University of Edinburgh.
36. Hobart, M. (2014). One Social Meaning – Two Languages: Sociophonetic Markers Indexing Identity in Bilinguals. Unpublished Master's Thesis. The University of Edinburgh.
37. Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*(1), 60–99.
38. Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America, 108*(3), 1252–1263.
39. Kachel, S., Simpson, A. P., & Steffens, M. C. (2018). "Do I sound straight?" - Acoustic correlates of actual and perceived sexual orientation and masculinity/femininity in men's speech. *Journal of Speech, Language, and Hearing Research*.
40. King, B. W. (2008). "Being Gay Guy, That is the Advantage": Queer Korean Language Learning and Identity Construction. *Journal of Language, Identity & Education, 7*(3–4), 230–252.
41. Labov, W. (1963). The Social Motivation of a Sound Change. *Word, 19*(3), 273-309.
42. Labov, W. (1966). The Social Stratification of English in New York City. Washington: Center for Applied Linguistics.
43. Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.
44. Levon, E. (2007). Sexuality in context: Variation and the sociolinguistic perception of identity. *Language in Society, 36*, 533–554.
45. Levon, E. (2014). Categories, stereotypes, and the linguistic perception of sexuality. *Language in Society*, *43*(5), 539-566.
46. Li, X. (2010). Sociolinguistic Variation in the Speech of Learners of Chinese as a Second Language. *Language Learning, 60*(2), 366–408.
47. Mack, S. (2010). A sociophonetic analysis of perception of sexual orientation in Puerto Rican Spanish. *Laboratory Phonology, 1*(1), 41–63.
48. Mack, S., & Munson, B. (2012). The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the "gay lisp" stereotype. *Journal of Phonetics, 40*(198–212).
49. Maegaard, M., & Pharao, N. (2016). /s/ variation and perceptions of male sexuality in Denmark. In E. Levon & R. B. Mendes (Eds.), *Language, Sexuality, and Power* (pp. 88–104). Oxford University Press.
50. Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America, 125*(6), 3962–3973.
51. Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech, 50*(1), 125–142.
52. Munson, B. (2010). Variation, implied pathology, social meaning, and the "gay lisp": a response to Van Borsel et al. (2009). *Journal of Communication Disorders, 43*, 1–5.

53. Munson, B., Jefferson, S. V, & McDonald, E. C. (2006). The influence of perceived sexual orientation on fricative identification. *Journal of the Acoustical Society of America, 119*(4), 2427–2437.
54. Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics, 34*(2), 202–240.
55. Papastergiadis, N. (2000). The turbulence of migration: Globalization, deterritorialization and hybridity. Malden, MA: Polity Press.
56. Pharao, N., & Maegaard, M. (2017). On the influence of coronal sibilants and stops on the perception of social meanings in Copenhagen Danish. *Linguistics, 55*(5), 1141–1167.
57. Pharao, N., Maegaard, M., Møller, J. S., & Kristiansen, T. (2014). Indexical meanings of [s+] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society, 43*, 1–31.
58. Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The influence of sexual orientation on vowel production. *Journal of the Acoustical Society of America, 166*(4), 1905–1908.
59. Podesva, R. J. (2006). Intonational variation and social meaning: Categorical and pho- netic aspects. *University of Pennsylvania Working Papers in Linguistics, 12*(2), 189–202.
60. Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in con- structing a persona. *Journal of Sociolinguistics, 11*(4), 478–504.
61. Podesva, R. J. (2011). Salience and the social meaning of declarative contours: Three case studies of gay professionals. *Journal of English Linguistics, 39*(3), 233–264.
62. Podesva, R. J., & Van Hofwegen, J. (2014). How Conservatism and Normative Gender Constrain Variation in Inland California: The Case of /s. *University of Pennsylvania Working Papers in Linguistics, 20*(2), 129.
63. Rácz, P., & Schepácz, A. (2013). The perception of high frequency sibilants in Hungarian male speech. *Acta Linguistica Hungarica, 60*(4), 457–468.
64. Regan, V. (1995). The Acquisition of sociolinguistic native speech norms: effects of a year abroad on L2 learners of French. In B. Freed (Ed.), *The Linguistic Impact of Study Abroad*. John Benjamins Publishing Co.
65. Rogers, H., & Smyth, R. (2003). Phonetic differences between gay- and straight-sounding male speakers of North American English. Proceedings of *The 15th International Congress of Phonetic Sciences*, 1855–1858.
66. Rosenfelder, I., Fruehwald, J., Evani, K., & Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite.
67. Russell, E. L. (2017). Style shifting and the phonetic performance of gay vs. straight: A case study from French. *Journal of Language and Sexuality, 6*(1), 128–176.
68. Sánchez, F. J., & Vilain, E. (2012). "Straight-acting gays": The relationship between masculine consciousness, anti-effeminacy, and negative gay identity. In *Archives of Sexual Behavior, 41*, 111–119).
69. Smyth, R., Jacobs, G., & Rogers, H. (2003). Male voices and perceived sexual orientation: An experimental and theoretical approach. *Language in Society, 32*(3), 329–350.
70. Smyth, R., & Rogers, H. (2002). Phonetics, gender, and sexual orientation. *The 2000 Meeting of the Canadian Linguistic Association*, Toronto, 299–311.
71. Smyth, R., & Rogers, H. (2008). Do gay-sounding men speak like women? *Toronto Working Papers in Linguistics, 27*.
72. Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research, 22*(3), 339–368.
73. Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In J. Cole & J. I. Hualde (Eds.), *Laboratory Phonology 9* (pp. 65–86). New York: Mouton de Gruyter.
74. Talbot, M. (2010). *Language and Gender* (2nd ed.). Cambridge, UK: Polity Press.
75. Thorpe, D. (2014). *Do I Sound Gay?* [Motion Picture]. United States: Sundance Selects.
76. Van Borsel, J., De Bruyn, E., Lefebvre, E., Sokoloff, A., De Ley, S., & Baudonck, N. (2009). The prevalence of lisping in gay men. *Journal of Communication Disorders*, *42*(2), 100-106.
77. Van Compernolle, R. A., & Williams, L. (2012). Reconceptualizing Sociolinguistic Competence as Mediated Action: Identity, Meaning-Making, Agency. *Modern Language Journal, 96*(2), 234–250.
78. Walker, A., García, C., Cortés, Y., & Campbell-Kibler, K. (2014). Comparing social meanings across listener and speaker groups: The indexical field of Spanish /s/. *Language Variation and Change, 26*(2), 169–189.
79. Wurr, A. (2003). Reading in a Second Language: A Reading Problem or a Language Problem? *Journal of College Reading and Learning, 32*(2), 157–169.
80. Zampini, M. L. (2008). L2 speech production research: Findings, issues, and advances. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 219–250). Amsterdam: John Benjamins Publishing.
81. Zhang, Q. (2005). A Chinese yuppie in Beijing: Phonological veriation in the construction of a new professional identity. *Language in Society, 34*(3), 431–466.
82. Zimman, L. (2015). Transmasculinity and the voice: Gender assignment, identity, and presentation. In T. M. Milani (Ed.), *Language and Masculinities: Performances, Intersections, Dislocations* (pp. 197–219). New York: Routledge.
83. Zimman, L. (2017). Variability in /s/ among transgender speakers: Evidence for a socially grounded account of gender and sibilants. *Linguistics, 55*(5), 993–1019.

© 2024 Zac Boyd. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

9. Appendix

Appendix 1: /s/ CoG: 95% CI (Reference Level - Gay/L1)

	Coefficient	Estimate		95% CI			exclude_0
	(Intercept)	6599	[	6110.47	7047.21	]	
T1*	**Sexuality Effect; Fre**	**-652.53**	**[**	**-1373.42**	**-39.99**	**]**	**TRUE**
T2*	**Sexuality Effect; Fre(Eng)**	**-893.84**	**[**	**-1607.09**	**-281.09**	**]**	**TRUE**
T3*	**Language Effect; (GayFre)**	**791.22**	**[**	**678.63**	**919.3**	**]**	**TRUE**
T4*	**Language Effect; (StrFre)**	**549.91**	**[**	**420.51**	**675.93**	**]**	**TRUE**
T5	Sexuality Effect; Ger	-424.94	[	-1017.14	127.51	]	FALSE
T6	Sexuality Effect; Ger(Eng)	-400.57	[	-1026.01	133.53	]	FALSE
T7*	**Language Effect; (GayGer)**	**230.69**	**[**	**132.49**	**312.43**	**]**	**TRUE**
T8*	**Language Effect; (StrGer)**	**255.07**	**[**	**157.3**	**347.26**	**]**	**TRUE**
T9	Nationality Effect; L1; Gay	187.19	[	-369.19	811.86	]	FALSE
T10	Nationality Effect; L2; Gay	373.34	[	-227.8	934.4	]	FALSE
T11	Nationality Effect; L1; Str	-414.77	[	-1084.82	224.46	]	FALSE
T12	Nationality Effect; L2; Str	-119.93	[	-758.98	536.07	]	FALSE

DOI: 10.2478/9788366675513-003

Assessing creaky voice quality for forensic purposes

Katharina Klug[1], Christin Kirchhübel[2], Paul Foulkes[1], Almut Braun[3], and Peter French[1,4]

[1]Department of Language and Linguistic Science, University of York, UK
[2]Soundscape Voice Evidence, Lancaster, UK
[3]Forensic Science Institute, Bundeskriminalamt, Wiesbaden, Germany
[4]J P French International, Zurich, Switzerland

katharinaklug@gmx.net, ck@soundscapevoice.com, paul.foulkes@york.ac.uk, almut.braun@bka.bund.de, peter.french@york.ac.uk

ABSTRACT

This study examines the multifaceted nature of creaky voice production (hereinafter CV), which results in diverse perception and acoustics. Therefore, CV categories need to be defined which are relevant for both perception and acoustics. Here, perceptually relevant CV categories are explored. A CV classification scheme is conceptualised and tested, which distinguishes between four CV modes (clean CV, harsh CV, breathy CV, and aperiodic creak) and acknowledges transitions into CV (e.g., modal voice – clean CV) as well as transitions between CV modes (e.g., harsh-breathy CV). The scheme is tested by four forensic speech scientists analysing spontaneous speech samples of six male English speakers. Results revealed that speakers vary with respect to the nature and frequency of preferred CV modes, i.e., CV space. We conclude that the nature of CV production may be useful to facilitate speaker discrimination.

Key words: creaky voice, perceptual analysis, forensic voice comparison, speaker-discriminating potential

1. Introduction

Creaky voice (henceforth CV) is frequently used in many languages as a short-term voice quality setting to mark various functions, e.g.: (1) contrastive phonation types, (2) lexical tone and register, (3) intonation, (4) vowel-initial glottalisation, and (5) coda-/t/ glottalisation (see [1] for a summary). Within forensic speech science, however, CV as a long-term setting is of particular interest. [2] reported speakers of American English not only to differ in their use of CV as a short-term or long-term setting, but also in their characteristic acoustic CV preferences. This variation is assumed to be based on anatomical, physiological, and habitual idiosyncrasies [3]. The aim of the present study is to determine if "persistent creakers" [4], who employ CV as a long-term voice quality setting, can be distinguished based on their particular CV production.

Forensic speech scientists analyse audio recordings related to legal (usually criminal) investigations. Frequently, the expert is asked to compare the speaker in an evidential recording with the recording of a known suspect, to address the likelihood of identity or non-identity. Analysing individual features like fundamental frequency, formants, rhythm, tempo, lexical choices, and voice quality, the level of similarity between the recordings is assessed, as well as the level of typicality in a defined reference population. Voice quality (henceforth VQ) is an important feature for the purpose of this task because of its potential power to discriminate between speakers [5]. We consider that this speaker-discriminatory power can be further exploited by taking advantage of the multifaceted nature of CV through analysing it on a finer scale.

CV production is typically characterised by short and thick vocal folds [6], with only the membranous part of the glottis being able to vibrate [7, 8]. [9] describes high adductive tension, high medial compression, and low longitudinal tension. The larynx is tensed [10] and typically elevated [8], which causes the adducted ventricular folds to couple with the vocal folds [11]. The airflow rate is low due to strong adduction [11, 12].

© 2024 Katharina Klug et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Various metaphors were used over the centuries to describe the perception of CV: [13, p. 32] refers to "a rapid series of taps, like a stick being run along a railing". [10, p. 231] compares it to "popping", "frying", "ticking" or "rasping" and [14, p. 2650] associates it with a "motorboat engine". All these analogies describe individual glottal pulses which are separately perceivable. Thus, we consider the presence of *distinct glottal pulses* to be the main characteristic of CV.

1.1. Previous work

Some studies acknowledged the multifaceted nature of CV and suggested potential ways to differentiate between types or modes. The following studies – ordered by year of publication – were chosen to be most relevant for forensic application and should be considered when developing a classification scheme, designed to meet the requirements for forensic application.

The Vocal Profile Analysis protocol (VPA) developed by Laver and colleagues [9, 15] has proven useful in analysing a speaker's VQ for forensic purposes and is used by forensic phoneticians in various adapted forms around the world [16]. The VPA assesses a speaker's VQ using a componential approach by capturing individual VQ aspects including vocal tract features (e.g., lip and larynx position), overall muscular tension (i.e., vocal tract and laryngeal tension), and phonation features (e.g., creaky and harsh voice). Each setting is assessed perceptually on a long-term basis using a 6-point scale, with a scalar degree of 4 to 6 being used for pathological voices. As for CV, according to [9] it is compatible with whispery voice, harsh voice, and falsetto and various combinations, e.g., harsh-whispery creaky voice.

Batliner et al. [17] developed the 'MÜSLI' scheme (Münchner Schema für Laryngalisierungs-Identifikation) which differentiates between glottalisation, damping, diplophonia (i.e., amplitude variations), subharmonics (i.e., frequency variations) and aperiodicity.

Slifka [18] studied CV production, which occurs at the end of utterances. It is characterised by irregular glottal pulses, abrupt glottal closing, and relatively rapid glottal opening. The vocal folds often lack proper glottal closure, resulting in increased airflow. [18] reported that incidences of this irregular CV mode are also found in non-final position.

Ishi and colleagues [19] highlighted the challenge to clearly distinguish between distinct and non-distinct glottal pulses, and thus the challenge to distinguish between presence and absence of 'vocal fry', a popular synonym for CV. They therefore introduced 'transitions' between vocal fry and modal voice, period-doubled voice, and harsh voice, recognising the phonation continuum. However, the authors did not study transitions systematically.

An example of a more systematic approach on the issue of transitions can be seen in Devaraj et al. [20]. Using synthesised vowels, the authors systematically varied fundamental frequency, open quotient, and amplitude quotient to assess the level of 'impulsiveness', i.e., the level of "temporal segregation of individual glottal pulses" [20, p. 2] on a 7-point-Likert scale. They demonstrated that different graduations of vocal fry impulsiveness can be perceived, which seemed to correlate primarily with f0 and somewhat with open quotient. Therefore, transitions into CV should also be considered within a CV classification scheme.

Moisik [8] described two different CV modes based on laryngeal height: *raised CV* and *lowered CV*. Although CV is typically produced with a raised larynx position, CV can also be produced with a lowered larynx which prevents vocal-ventricular fold coupling and biases "phonation towards breathiness" [8, p. 211]. The longitudinal tension of the glottis and the aryepiglottic constriction above the glottis are considerably less than in raised CV.

Keating et al. [21] described low f0 and irregular f0 to be sufficient to generate CV perception. They list glottal constriction as a further key property, which, on its own however, is not sufficient to generate creaky perception. All three key properties can occur in combination with each other as prototypical creak (low f0 + irregular f0 + constricted glottis), spread glottis creak (low f0 + irregular f0), and vocal fry (low f0 + constricted glottis). Furthermore, low f0 can occur together with multiple pulses.

Each of the classification approaches summarised above provided valuable perspectives on the multifaceted nature of CV and was taken into account when developing the proposed CV classification scheme.

1.2. Scope and Research Questions

The scope of the study is to explore the possibility of improving the analysis of CV by implementing a workable CV classification scheme with perceptually relevant categories. The approach is based on perceptual analysis, with visual signal characteristics used to complement the perceptual categories. Given the project's scope, the following research questions are posed:

RQ1
CV assessment: Is it possible to assess the *nature of CV production* using the proposed CV classification scheme? How *consistently do analysts* perform?

RQ2
CV production: Can *speakers be distinguished* by the nature of their CV production?

2. Method and Data

2.1. Method

The current study considers the presence of *distinct glottal pulses* as the main criterion of CV perception. Distinct glottal pulses can have two causes: (1) damping effects, i.e., amplitude attenuation, or (2) aperiodicity. (1) The vocal tract in CV is described to be highly damped between glottal excitations [22, 23]. Possible causes for these damping effects are coupling effects between the vocal folds and the ventricular folds [24], a long glottal closure phase [25], or a slow glottal opening phase [26]. Damping effects cause energy losses [19], thus enabling the perception of individual pulses. (2) The perception of distinct glottal pulses, however, can also be generated by *aperiodic glottal pulses*, which often appear as long pulses, i.e., low pitch, typically listed as a CV characteristic [e.g., 27, 28]. Human listeners are very sensitive to aperiodicity. [29] revealed that the presence of only one or two slightly aperiodic pulses are sufficient for CV perception. These two causes for CV perception, i.e., damping effects and aperiodicity, form the core of the proposed CV classification scheme.

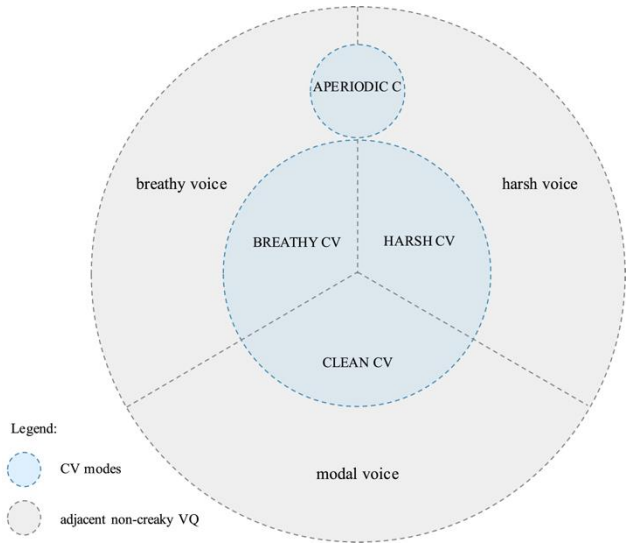


Figure 1: *proposed CV classification scheme illustrating CV modes and adjacent non-creaky VQs.*

The CV scheme (Figure 1) is divided into a blue inner section surrounded by a grey outer section. The blue section displays the presence of distinct glottal pulses and thus CV, while the grey section displays adjacent non-creaky VQs which lack distinct glottal pulses. CV is separated into four CV modes. Three of them, i.e., *clean*, *harsh,* and *breathy* CV, produce CV by amplitude damping effects. *Aperiodic creak* (henceforth aperiodic C) is characterised by aperiodically spaced glottal pulses and thus segregated from the amplitude damped CV modes in the scheme. The CV modes are ordered according to perceptual proximity and/or distance. Aperiodic C is perceptually furthest from clean CV but may slightly overlap with harsh CV and breathy CV. [6] reported that aperiodic noisy voices perceptually overlap slightly with period-doubled voices. Aperiodic C results in a noisy and rough quality, comparable to harsh and breathy CV. Therefore, aperiodic C perceptually borders on harsh-breathy CV.

Within the grey section, phonation types which are most closely related to CV are displayed, i.e., modal voice, harsh voice, and breathy voice. We follow [19] and [20] in acknowledging the continuity of glottal pulse distinctiveness and thus between *CV and non-CV* and between *CV modes*, as indicated by the dashed lines.

Examples of waveforms and broad-band spectrograms (window length 0.03s) for each CV mode are shown below (Figures 2-5). They were selected based on 100% agreement between all analysts from the main study. In the absence of a clear example of breathy CV, the transition harsh-breathy CV is illustrated instead. Dashed lines mark the CV section.

2.1.1. Clean CV

Clean CV is characterised by distinct glottal pulses and the absence of additional perceptual characteristics. Clean CV can appear tense. The larynx is typically elevated [8]. Figure 2 clearly shows the damping between the glottal pulses. In the waveform, excitation amplitudes are clearly separated from each other by damped amplitudes. The same phenomenon can be seen in the spectrogram below: the dark vertical lines are followed by white lines, indicating the absence of energy. The vertical distinct lines in the spectrogram, which extend over the entire frequency range indicate the distinct glottal pulses [9].

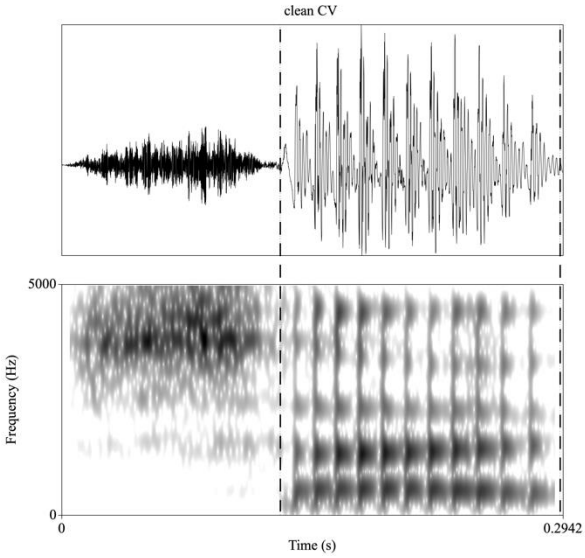


Figure 2: *example for clean CV (speaker 004 <so>).*

2.1.2. Harsh CV

With *harsh CV*, the perception of distinct glottal pulses occurs simultaneously with the perception of harshness, i.e., roughness. Harsh voice quality is produced by coupling the vocal folds with either the ventricular folds or the

© 2024 Katharina Klug et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

aryepiglottic folds [30]. This coupling effect results in amplitude modulations, subharmonics, or chaos [31]. Acoustically, subharmonics and amplitude modulations result in multiple pulses. Perceptually, either two simultaneous pitches [32] or an indeterminate pitch is to be expected. Chaos is perceived as particularly harsh as the vocal folds oscillate highly irregularly, resulting in harmonic smearing and broadband noise [31]. The waveform of harsh CV in Figure 3 shows regular amplitude modulations resulting in two repetitive patterns. These amplitude modulations are also evident in the spectrogram as weak vertical lines between the strong vertical lines.

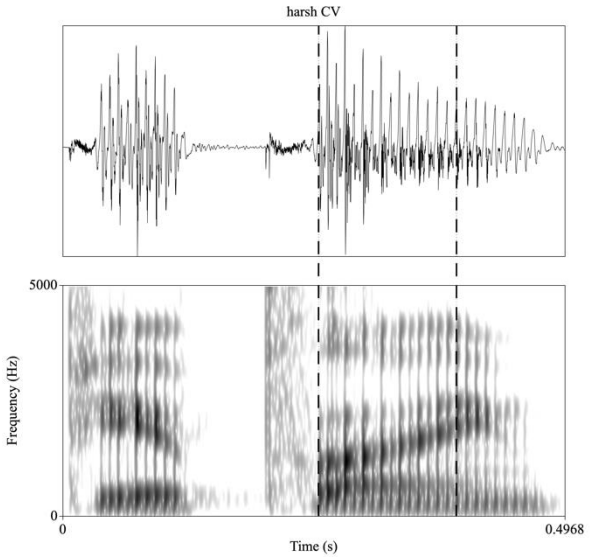


Figure 3: *example for harsh CV (speaker 006 <be applying>).*

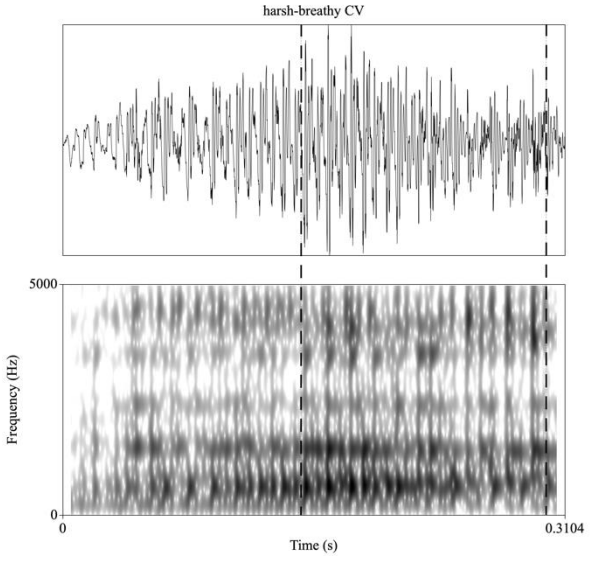


Figure 4: *example for harsh-breathy CV (speaker 005 <uh>).*

2.1.3. Breathy CV

Breathy CV is generated when distinct glottal pulses occur simultaneously with high frequency noise, and thus the impression of breathy/whispery voice quality that occurs together with CV. According to [33] the high frequency noise is particularly apparent in the frequency region around F3. The larynx is lowered [8]. Previous studies use the following terminologies: *whispery CV* [9], *breathy-laryngealized mode of vibration* [33], *lowered/lax creaky voice* [8], or *spread glottis creak* [21]. The example in Figure 4 displays harsh-breathy CV, the transitional mode between harsh CV and breathy CV. Here, distinct glottal pulses occur simultaneously with high frequency noise and chaos. This becomes apparent by the less periodic waveform with irregular amplitude modulations. The spectrogram reveals less distinct vertical lines with considerable energy in between and less prominent formants, all of which indicates the presence of noise.

2.1.4. Aperiodic creak

Aperiodic Creak is characterised by distinct glottal pulses resulting from aperiodic vocal fold oscillation rather than amplitude damping alone. Aperiodicity inherently prevents a clear pitch perception, and thus tends to be perceived as "noisy". The lack of pitch is also reflected in the term *creak* rather than *creaky voice*. In contrast to the three CV modes which are based on long-term amplitude damping effects, *aperiodic C* occurs as a short-term setting. As stated above, human listeners are very sensitive to even slightly aperiodic pulses [29], and thus very susceptible to aperiodic C. An example is shown in Figure 5. The aperiodicity becomes apparent through the irregular main excitation peaks in the waveform, which are also clearly visible as irregular occurring dark vertical lines in the spectrogram.

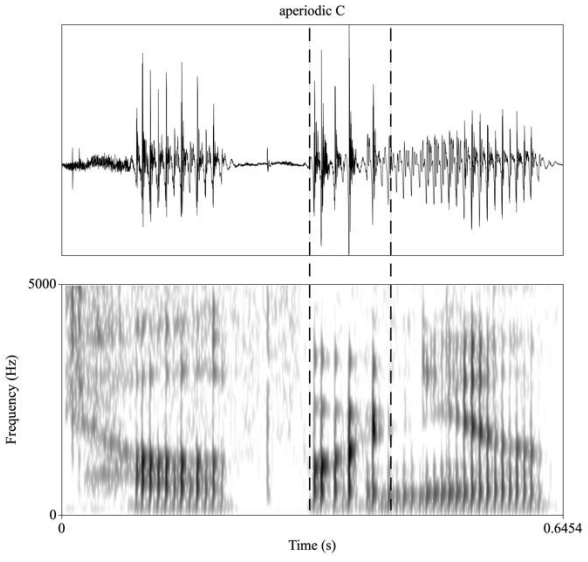


Figure 5: *example for aperiodic C (speaker 002 <California>).*

2.1.5. Adjacent voice qualities

The proposed VQs adjacent to CV are *modal voice*, *harsh voice*, and *breathy voice*. As described by [19], adjacent VQs are not clearly delineated from CV, either in terms of production or perception. The proposed adjacent VQs are hypothesised to share a continuum with CV. This continuum may result in analyst-specific thresholds when assessing presence or absence of distinct glottal pulses and could therefore affect the consistency between (and within) analysts when applying the CV classification scheme.

2.2. Data

A corpus of six speakers was compiled. The samples contained spontaneous speech of male English speakers, provided at 44.1 kHz sampling frequency and a bit depth of 16-bit. Each speaker was represented by three samples that were taken from the same recording, each about 30 seconds long. This resulted in a total of 18 samples. The recordings were obtained from five corpora of conversational speech, all involving UK or US English: WYRED [34], SpeechBox Allsstar [35], The Life Scientific [36], Northern Englishes [37] and Deceptive Speech [38]. The speakers were selected by the lead author to represent a wide range of CVs, including speakers who straddle the line between creaky and non-creaky VQ. Praat TextGrid files [39] were provided, specifying the syllables to be analysed. Only stressed syllables were chosen to exclude syllables phonated with low subglottal pressure. Filled particles *uh* and *um* were included too, as they are prone to creaky phonation [40]. 20 syllables were labelled within each sample, resulting in 60 syllables per speaker, 360 syllables per analyst and 1440 syllables in total.

2.3. Procedure

To test the proposed CV classification scheme, a focus group was formed, consisting of four forensic speech scientists (the first to fourth authors), selected for their interest in VQ and CV in particular. All focus group participants had similar training in perceptual VQ analysis, based on the application of the Vocal Profile Analysis by Laver and colleagues (15) and regularly analysed VQ for forensic casework and/or VQ research. The testing procedure involved an alternation of individual analytical listening tasks followed by in-depth online discussion sessions to address challenges and evaluate results. For the pilot test, training material was provided to the focus group participants to ensure familiarisation with the proposed CV classification scheme. This included (1) a summary with basic perceptual characteristics of the three amplitude damped CV modes, i.e., clean CV, harsh CV, and breathy CV, and (2) three short sample recordings selected by the first author to represent each of the CV modes. Aperiodic C was not included as it was only introduced after the pilot test. In the subsequent online discussion session, difficulties were discussed which provided helpful insight for the design of the main study.

In this paper the main study is discussed. Within the main study the participants were instructed to follow a strict assessment method, dividing the classification process into the following steps:

1. DistGloPuls: Is the syllable produced with distinct glottal pulses?
    a. yes
    b. no → NO CV
2. OVQ [Other Voice Quality]: If the syllable is not produced with DistGloPuls, please state how you would characterise the syllable's main VQ instead. (e.g., modal voice, harsh voice, breathy voice)
3. f0A/AmD: If the syllable is produced with DistGloPuls, are these generated due to *frequency aperiodicity* (f0A) or due to *amplitude damping* (AmD)?
    a. f0A → APERIODIC CV
    b. AmD
4. FurChar: If the DistGloPuls are due to AmD, are there any *further characteristics*, i.e., *multiple pulses* (MulPul), *high frequency noise* (Noi)?
    a. no → CLEAN CV
    b. MulPul → HARSH CV
    c. Noi→ BREATHY CV
5. CVmode: Depending on the entries of the tiers above, the respective *CV mode* can be determined (see Table 1). As VQ is continuous, 'in-between' transitional modes are also possible (e.g., harsh-clean CV, clean CV-modal voice).
6. Notes: Any *comments* or *questions* can be entered here.

The flowchart in Table 1 illustrates the stepwise classification process. This fine-grained procedure was designed to test the workability of the CV classification scheme in the first instance. Thus, a highly controlled context was chosen where samples were 'pre-processed', i.e., analysts were directed to examine isolated syllables specifically chosen by the experimenter. If the scheme proves feasible in this context, then there are strong arguments for further exploring its feasibility in a less controlled (and forensically more realistic) environment, i.e., by applying it to samples of speech which are not pre-processed.

Calibration for the main study was carried out by providing sample recordings for each proposed CV mode together with a description of the expected characteristics visible in waveforms, broad-band and narrow-band spectrograms and FFT spectral slices. Analysis was based primarily on perceptual cues. Visual signal inspection could be used to corroborate perception.

© 2024 Katharina Klug et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Distinct glottal pulses	Absent	NO CV			
	Present	Frequency aperiodicity	APERIODIC CV		
		Amplitude damping	Further characteristics	Absent	CLEAN CV
				Present	Multiple pulses	HARSH CV
					High frequency noise	BREATHY CV

Table 1: *flowchart for assessing CV modes.*

2.4. Data analysis

For visualisation purposes, the syllable classifications provided by the four analysts (e.g., harsh CV) were transferred into x-y-coordinates. Using Microsoft Excel the CV classification scheme was placed on a grid so that all syllable classifications made for each speaker could be plotted into the scheme. By using a bubble chart, the frequency of occurrences of each classification was represented through bubble size. Colour was used to represent the ratings of individual analysts. Figure 7 shows the resulting bubble charts, which allow for comparisons between analysts within a chart, and between speakers across charts.

Quantitative analysis was performed by calculating the frequency with which each CV mode was assigned per speaker, both per analyst and across analysts (Table 2).

Agreement between analysts was assessed in two ways: (1) Agreement in classifying presence or absence of distinct glottal pulses (DistGloPuls) was calculated using Gwet's agreement coefficient (AC2). (2) Agreement in classifying the phonation mode of individual syllables was analysed using a distance measure calculated for each analyst pair.

3. Results and Discussion

The results are presented in terms of (1) CV assessment, and (2) CV production. Table 2 and Figures 6 and 7 illustrate the results.

Table 2 lists the *preferred CV modes* for each speaker. Speaker *001*, for example, preferred three CV modes – clean CV, harsh CV, and clean-harsh CV. The percentages illustrate the proportions with which particular CV modes were assigned. Thus, 41% of speaker *001's* syllables were rated as clean CV when averaged across the four analysts. To be classified as a *preferred CV mode,* two conditions must be met: (1) the CV mode occurs in at least 10% of all CV ratings *across all analysts* (see *% across analysts*, 1st row in Table 2), and (2) the same CV mode occurs in at least 10% of all CV ratings *by at least two analysts* to ignore analyst-specific outliers (see *% per analysts*, 2nd row in Table 2, percentages of individual analysts separated by slashes).

Figure 6 displays the same information as in Table 2 *(% across analysts*, 1st row) visually within the CV classification scheme. Each speaker is represented by a different colour. Only the part of the scheme covering CV is shown, while adjacent VQs are ignored.

The bubble charts in Figure 7 illustrate the ratings of *each analyst* for each of the six speakers separately. Here, different colours represent different analysts.

3.1. CV assessment

To assess the level of consistency between analyst, we need to differentiate between two aspects: (1) determination if CV is present, i.e., whether a syllable can be characterised by distinct glottal pulses (DistGloPuls), and (2) assessment of phonation mode.

3.1.1. DistGloPuls

The threshold between distinct and non-distinct glottal pulse oscillation is continuous. Therefore, we distinguish between three perceptual gradations to specify DistGloPuls: presence, absence, and inconclusiveness. Inconclusiveness of DistGloPuls was inferred when analysts indicated uncertainty when assessing DistGloPuls by either adding a question mark or indicating "slightly" or "borderline" in the comment box.

Analyst agreement on DistGloPuls was calculated using the

Gwet's agreement coefficient (AC2) for more than two analysts from the *irrCAC* package [41] in Rstudio [42]. The calculation was based on the ordinal raw ratings for each syllable and each analyst. Percent agreement reached AC2=0.56 for all analysts, indicating *moderate* agreement [according to 43], while individual analyst pairs ranged between AC2=0.48 (*moderate*) and AC2=0.61 (*substantial*). Note that this scale indicates that the degree of agreement was well above chance level. Overall, the analysts in the current study seem to be in broad agreement when assessing the presence of distinct glottal pulses.

3.1.2. Phonation mode

The agreement between analysts when assessing the phonation mode of each syllable was calculated for each analyst pair by determining the perceptual distances between each syllable's ratings. Distances were determined as follows: (1) the distance between two amplitude damped CV modes (e.g., clean CV and harsh CV) was defined as 1 distance. The non-creaky VQs were treated similarly, e.g., modal voice and breathy voice were considered as 1 distance point apart. (2) As aperiodic C is perceptually further away from clean CV than it is from harsh and breathy CV, disagreement between aperiodic C and clean CV was classified to be greater (distance 2) than between aperiodic C and harsh or breathy CV (distance 1). (3) When analysts differed in their classification only by an adjacent transitional mode (e.g., clean CV vs. clean CV-modal voice) the distance between both classifications was rated to be 0.5. (4) The largest possible perceptual distance within the proposed CV classification scheme was considered between aperiodic C and modal voice and rated as 3 distances.

For four speakers, namely *001*, *002*, *003* and *005*, all analyst pairs classified the same syllables with an average distance of 0.5, which corresponds to distance category (3) above, i.e., the distance between clean CV and clean-harsh CV (see *% per analyst* of analysts 1 and 2 for speaker *001* in Table 2), or clean CV and clean CV-modal voice (speaker *002* in Table 2). For these four speakers, the individual analyst pairs ranged between 0.05 distances, which is almost perfect, to 0.9.

For speakers *004* and *006* analysts showed lower values of agreement when classifying the phonation mode. Across all analyst pairs the averaged distance was 0.9 with individual analyst pairs ranging between 0.4 and 1.3 distances.

Across all speakers each analyst pair showed similar agreement performances with distance values ranging between 0.6 and 0.7 distances. Consequently, no analyst pair was in greater agreement across all speakers than all other analyst pairs.

3.1.3. Further insights

The focus group discussion sessions revealed that apparent disagreement between analysts regarding DistGloPuls and phonation mode did usually not stem from differences in perception, but rather from the way in which the perception was classified. Often this was due to analyst-specific thresholds.

	001	*002*	*003*	*004*	*005*	*006*
Preferred CV mode % across analysts % per analyst	clean CV 41 4/74/41/46	clean CV 32 13/57/33/25	clean CV 66 64/79/58/63	clean CV 36 43/68/24/14		clean CV 14 17/13/13/14
	harsh CV 11 0/0/32/14				harsh CV 19 14/30/38/3	harsh CV 21 38/33/17/6
		aperiodic C 27 6/35/43/25			aperiodic C 31 19/42/43/26	aperiodic C 33 17/46/50/23
	clean-harsh CV 29 68/9/18/20				harsh-breathy CV 25 54/0/10/26	
		clean CV-modal voice 29 55/5/23/33	clean CV-modal voice 11 18/5/6/13		harsh CV-harsh voice 16 5/6/10/40	clean CV-modal voice 14 25/8/13/11

Table 2: *preferred CV modes per speaker. Numbers indicate the percentage with which the classification was assigned from all CV ratings made, both across all analysts and per analyst (separated by slashes).*

© 2024 Katharina Klug et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

3.2. CV production

It can be seen from Figures 6 and 7, that breathy CV featured much less frequently compared to the other CV modes. Indeed, it did not quality as a preferred CV mode for any of the speakers. This may suggest that breathy CV is more unusual compared to the other CV modes.

In contrast, clean CV featured much more extensively among the speakers analysed. Although clean CV appears to be a more frequent and therefore a more typical CV mode, this does not mean that two speakers cannot be distinguished from each other just because they use clean CV as their preferred CV mode. This is because, generally, speakers have more than one preferred CV mode and that combination of preferred CV modes differs between speakers. For example, while speakers *001* and *002* both have clean CV as a preferred CV mode, they differ in terms of additional preferred CV modes, with speaker *001* using harsh CV and speaker *002* preferring aperiodic C. In fact, it is only speaker *003* (grey in Figure 6), whose preferred CV modes only revolve around clean CV. Whether a lack of diversity of CV modes within a speaker can be considered more unusual than a speaker displaying a variety of CV modes warrants further investigation.

Speaker *005* (purple) was the only speaker who did not prefer clean CV at all but instead spanned the upper-right corner of CV modes, i.e., aperiodic C and all modes which include some degree of harshness. This may suggest that the absence of clean CV is more unusual and therefore has a higher speaker idiosyncratic value.

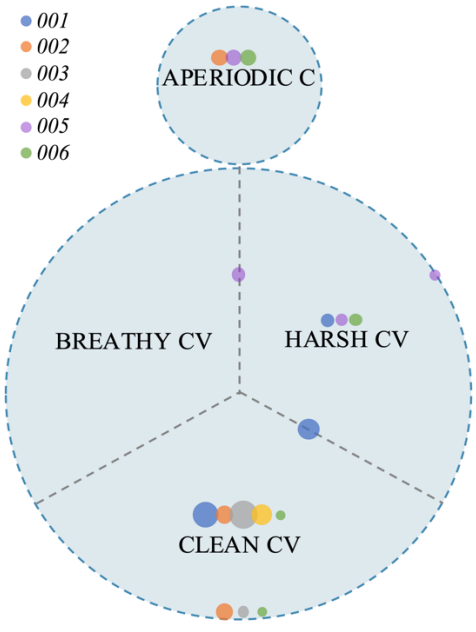


Figure 6: *preferred CV modes per speaker across all analysts.*

Some pairs of speakers, such as *003* and *005*, can be clearly distinguished from each other. Other speakers' CV spaces overlap to a greater degree, e.g., *003* vs. *004* (grey vs. yellow), *002* vs. *006* (orange vs. green), and *001* vs. *004* (blue vs. yellow). Despite the occurrence of overlap, none of the six speakers share the same CV space, i.e., there is no complete overlap between any of the speakers with respect to the nature and frequency of their preferred CV modes.

Figure 7 supports the results by also including the syllables that were classified as non-creaky but directly bordered the speaker's CV spaces. This is most evident in speakers *002* and *005*.

4. Conclusion

Based on spontaneous speech samples of six male English speakers, the present study explored the possibility to assess the multifaceted nature of CV further by subdividing it into four main CV modes, i.e., clean CV, harsh CV, breathy CV, and aperiodic C. Explicit recognition was given to transitional modes such as harsh-breathy CV, or harsh CV-harsh voice. Looking back at the research questions, we conclude as follows.

RQ1 CV assessment: Is it possible to assess the *nature of CV production* using the proposed CV classification scheme? How *consistently do analysts* perform?

Analysts showed moderate agreement when assessing the presence/absence of distinct glottal pulses, which was defined as the main characteristic of CV in the current study. When assessing the nature of the CV modes, the analysts showed strong agreement on four out of the six speakers using the proposed CV classification scheme. Discussion sessions revealed that disagreement between analysts usually did not stem from different perceptions; instead, often the same perception was just differently classified. Particular attention should therefore be paid to the calibration of the analysts, e.g., by providing more detailed training materials or conducting regular calibration sessions.

RQ2 CV production: Can *speakers be distinguished* by the nature of their CV production?

Overall, the speakers in the current study differed in the nature with which they produced CV. While some of the speakers could be more easily distinguished than others, even those with a higher level of similarity still did not fully overlap in their preferred CV space. Further, some CV modes or combination of CV modes appear to be more unusual than others. The results suggest that the fine-grained assessment of CV in a forensic voice comparison could help to discriminate between speakers. Thus, the speaker-discriminating potential of CV production should be further explored by applying the proposed CV classification scheme to a larger set of speakers. Analysing a larger sample would shed more light on the typicality of CV modes and CV spaces. Further testing would also reveal whether the proposed CV classification scheme captured the most important dimensions of CV production or whether revision is needed.

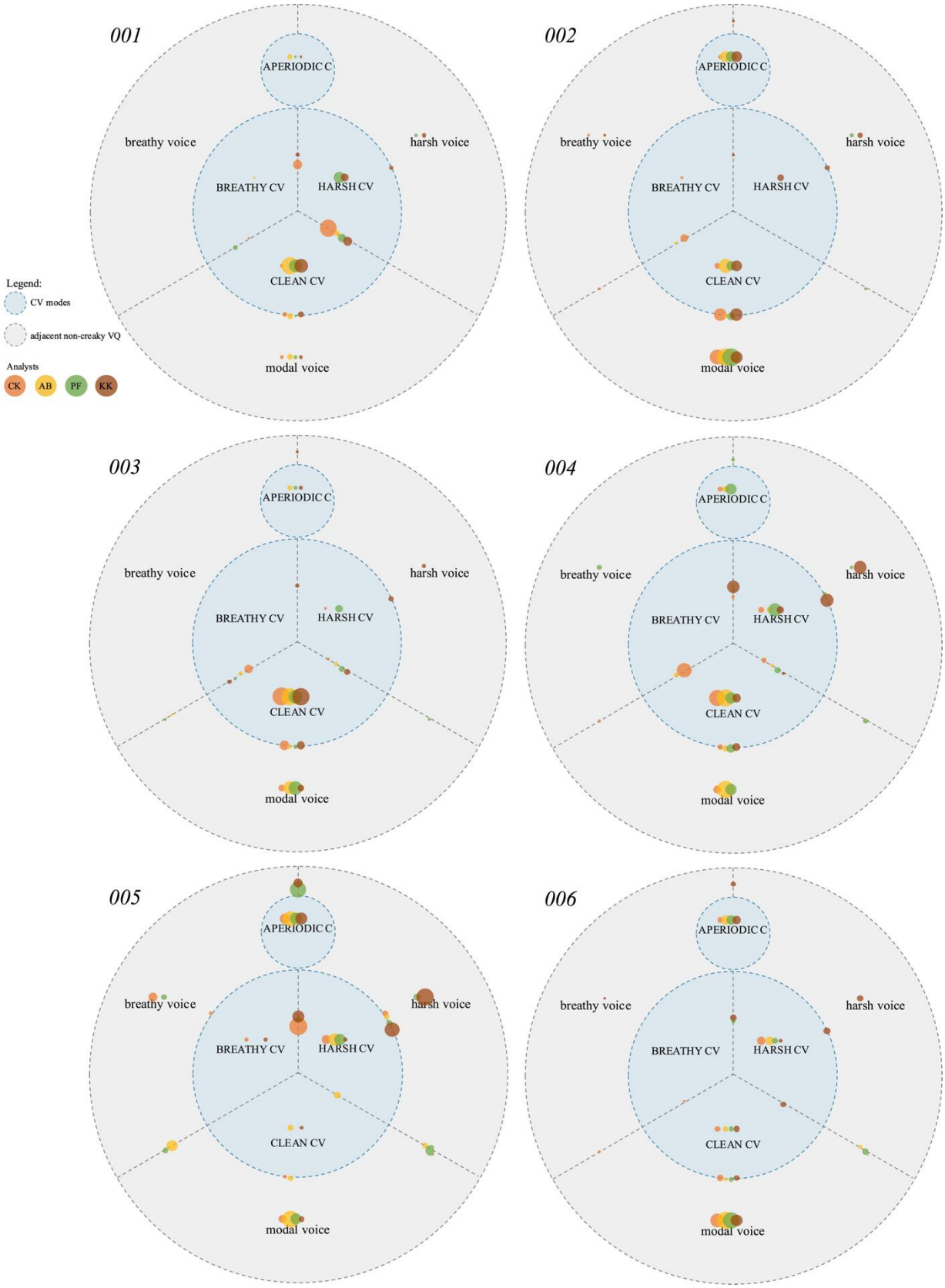


Figure 7: *ratings of all labelled syllables per speaker and per analysts.*

© 2024 Katharina Klug et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Successful implementation of the method in casework practice requires the provision of training material that enables calibration and thus harmonises analyst-specific threshold values. The training material should include sample recordings from various speakers which exemplify (1) thresholds for distinct glottal pulse presence, and (2) highlight expected perceptual characteristics of potential CV modes. In addition, visual cues from waveforms and spectrograms should be made explicit which may be used to corroborate perception. The analysis process also requires simplification, as it is too time-consuming for a casework application – comparing one speech sample with another speech sample may already take up to 15 hours.

A future aim could be to determine the nature of CV not just based on perceptual-visual analysis but to also use acoustic information to corroborate perception. A combined auditory-acoustic approach adopts the "complementary strengths of the two approaches" [44, p. 765]. To incorporate acoustic analysis, it needs to be discovered which acoustic measurements correlate with the perceptually relevant CV categories. Approaches such as [20] and [21] provide suggestions for relevant acoustic parameters which could be explored in this respect. Implementing an auditory-acoustic approach for assessing laryngeal VQ such as CV may further increase transparency and comparability of these assessments.

5. Acknowledgments

We would like to thank Patricia Keating, Jessica Wormald and one anonymous reviewer for valuable comments on previous versions of the paper.

6. References

1. Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, 94: Article 101155.
2. Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4), 423-444.
3. Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29(4), 407-429.
4. Henton, C., & Bladon, A. (1988). Creak as a sociophonetic marker. In Hyman, L. M., & C. N. Li (eds.), *Language, speech, and mind: Studies in honour of Victoria A. Fromkin* (pp. 3-29). London, New York: Routledge.
5. Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18(2), 293-307.
6. Gerratt, B., & Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(4), 365-381.
7. Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. University of Chicago Press.
8. Moisik, S. R. (2013). *The epilarynx in speech*. [Doctoral dissertation, University of Victoria].
9. Laver, J. (2009). *The phonetic description of voice quality*. Cambridge University Press.
10. Moore, P., & Von Leden, H. (1958). Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica et Logopaedica*, 10(4), 205-238.
11. Hollien, H. (1974). On vocal registers. *Journal of Phonetics*, 2(2), 125-143.
12. Gobl, C., & Ní Chasaide, A. (2010). Voice source variation and its communicative functions. In Hardcastle, W. J., J. Laver, & F. E. Gibbon (eds.), *The handbook of phonetic sciences* (pp. 378-423). Wiley-Blackwell.
13. Catford, J. C. (1964). Phonation types: The classification of some laryngeal components of speech production. In Abercrombie, D., D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (eds.), *In honour of Daniel Jones: papers contributed on the occasion of his eightieth birthday 12 September 1961* (pp. 26-37). London: Longmans, Green & Co.
14. Blomgren, M., Chen, Y., Ng, M. L., & Gilbert, H. R. (1998). Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *Journal of the Acoustical Society of America*, 103(5), 2649-2658.
15. Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. M. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*, 14, 139-155.
16. San Segundo, E. (2021). International survey on voice quality: Forensic practitioners versus voice therapists. *Estudios de fonética experimental*, 9-34.
17. Batliner, A., Burger, S., Johne, B., & Kießling, A. (1993). MUSLI: a classification scheme for laryngealizations. *Proc ESCA Workshop on prosody, Lund, Sweden*, 176-179.
18. Slifka, J. (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice*, 20(2), 171-186.
19. Ishi, C. T., Sakakibara, K. I., Ishiguro, H., & Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), 47-56.
20. Devaraj, V., Roesner, I., Wendt, F., Schoentgen, J., & Aichinger, P. (2023). Auditory Perception of Impulsiveness and Tonality in Vocal Fry. *Applied Sciences*, 13(7): Article 4186.
21. Keating, P. A., Garellek, M., Kreiman, J., Chai, Y. (2023). Acoustic properties of subtypes of creaky voice. *Journal of the Acoustical Society of America*, 153(3): Article 297.
22. Wendahl, R. W., Moore, G. P., & Hollien, H. (1963). Comments on vocal fry. *Folia Phoniatrica et Logopaedica*, 15(4), 251-255.
23. Coleman, R. F. (1963). Decay characteristics of vocal fry. *Folia Phoniatrica et Logopaedica*, 15(4), 256-63.

24. Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). *Voice quality: the laryngeal articulator model*. Cambridge University Press.
25. Fant, G. (1979). Glottal source and excitation analysis. *STL-Quarterly Progress and Status Report*, 1, 85-107.
26. Murry, T. (1971). Subglottal pressure and airflow measures during vocal fry phonation. *Journal of Speech and Hearing Research*, 14(3), 544-551.
27. Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. *Proc 19th International Congress of Phonetic Sciences, Melbourne, Australia,* 532-536.
28. White, H., Penney, J., Gibson, A., Szakay, A., & Cox, F. (2021). Optimizing an automatic creaky voice detection method for Australian English speaking females. *Proc 22nd Interspeech, Brno, Czech Republic,* 1384-1388.
29. Docherty, G. J., Foulkes, P., Milroy, J., Milroy, L., & Walshaw, D. (1997). Descriptive adequacy in phonology: A variationist perspective. *Journal of Linguistics*, 33(2), 275-310.
30. Moisik, S. R. (2013). Harsh voice quality and its association with blackness in popular American media. *Phonetica*, 69(4), 193-215.
31. Anikin, A., Pisanski, K., Massenet, M., & Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proc Royal Society B*, 288(1954): Article 20210872.
32. Kramer, E., Linder, R., & Schönweiler, R. (2013). A study of subharmonics in connected speech material. *Journal of Voice*, 27(1), 29-38.
33. Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2), 820-857.
34. Gold, E., Ross, S., & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proc 19th Interspeech. Hyderabad, India,* 2748-2752.
35. Bradlow, A.R. (n.d.) ALLSSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings. [Data Collection].
36. Al-Khalili, J. (Host). (2022, Oct 18). A passion for fruit flies [Audio podcast]. In *The life scientific*. BBC radio 4.
37. Haddican, W., & Foulkes, P. (2017). A comparative study of language change in Northern Englishes. [Data Collection]. Colchester, Essex: ESRC.
38. Kirchhübel, C. (2013). *The acoustic and temporal characteristics of deceptive speech.* [Doctoral dissertation, University of York].
39. Boersma, P. & Weenink, D. (2023). *Praat: doing phonetics by computer* (Version 6.3.08) [Computer software].
40. Muhlack, B., Trouvain, J., & Jessen, M. (2023). Distributional and Acoustic Characteristics of Filler Particles in German with Consideration of Forensic-Phonetic Aspects. *Languages*, 8(2): Article 100.
41. Gwet, K. L. (2019). irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). R package (Version 1.0) [Computer Software]
42. R Core Team (2021). *R: A language and environment for statistical computing* (Version 4.1.0) [Computer software].
43. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159-174.
44. Nolan, F. (1997). Speaker recognition and forensic phonetics, In Hardcastle, W. J. & J. Laver (eds), *The handbook of phonetic sciences* (pp. 744-767). Blackwell.

© 2024 Katharina Klug et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-004

Age and gender variation in Scottish voice quality beyond creak: a multi-measure approach

Joe Pearce[1]

[1]University of Glasgow, UK
j.pearce.1@research.gla.ac.uk

ABSTRACT

Research on Scottish accents using Vocal Profile Analysis [1] suggests that voice qualities such as whispery, tense, creaky, and harsh voice [2, 3, 4, 5] play a role in sociophonetic variation in voice quality in Scottish accents. However, sociophonetic variation in Scottish voice quality has never been considered in a large-scale acoustic study. Having separated creaky voice as described in [6] using an automated f0-based approach [7], here I investigate variation in non-creaky voice quality according to age, gender, and area in non-creaky stretches of sonorant sounds taken from spontaneous speech of 95 speakers of Glasgow, Lothian and Insular Scots from SCOSYA [8]. As acoustic measurements of voice quality can be difficult to interpret in isolation because of the multidimensional nature of the voice [9-11], here I combine H1*−H2* and CPP in a multi-measure approach. Female speakers showed evidence of a tenser voice quality than male speakers, while younger speakers displayed a more modal-like quality than older speakers. These findings give evidence to support the idea that female speakers are not universally breathier than male speakers and emphasize the importance of considering the voice as a multidimensional entity in sociophonetic research.

Key words: phonation, voice quality, sociophonetics, corpus, Scottish English, Scots, gender, ageing

1. Introduction

Auditory-perceptual research using Vocal Profile Analysis (VPA) [1] has revealed that whispery voice is prevalent in Scottish accents [2, 3] and that whispery, creaky, tense [3], and harsh voice [4] vary by age, gender, social class and region [2, 3, 4]. Recent advances in acoustic analysis of voice quality, such as automatic detection of creaky voice [7] and multi-measure acoustic analysis using psychoacoustically-validated measures [9-11], mean that this variation may now be captured using acoustic methods. Here, I consider the sociophonetic patterning of voice quality in Scottish accents in the Scots Syntax Atlas corpus (SCOSYA) [8] by age, gender, and area using acoustic methods, considering how a multi-pronged, multi-measure approach to acoustic analysis can allow sociophonetic variation in voice quality to be interpreted with reference to auditory quality.

1.1. Previous work on Scottish voice quality

Existing research on Scottish voice quality has used VPA, an auditory-perceptual approach to analysing voice quality in terms of its laryngeal and supra-laryngeal components developed by Laver and colleagues [1]. I restrict my focus here to the laryngeal voice quality settings, which in VPA are evaluated on scalar degrees with reference to modal voice. Being developed alongside Laver's systematic description of voice quality [12, 13], this framework has the advantage of being directly interpretable in terms of descriptive phonetic labels for voice quality such as modal, breathy, creaky, whispery, and harsh voice.

Using VPA, Esling [4] considered sociophonetic variation in 32 adult male speakers and 18 boys from Edinburgh divided into three groups of socio-economic status, and found that harsh voice predominated in speakers with lower socio-economic status, while the higher socio-economic status group used more creaky voice, and some degree of whispery voice was present across groups. Beck [2] considered voice quality among 50 speakers without voice disorders resident in Edinburgh, mostly with south-east Scottish accents, as part of a study on organic variation in voice quality. She found that whispery voice was prevalent in the sample: all speakers were rated at scalar degree 2 or higher for whispery voice. Furthermore, male speakers were more creaky than female speakers, being rated on average a scalar degree higher than female speakers in terms of creaky voice. Stuart-Smith [3] looked at voice quality in 32 speakers from Glasgow stratified by age, gender, and social class in word lists,

© 2024 Joe Pearce. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

read speech, and conversational speech. She found tense whispery voice characterised Glaswegian laryngeal voice quality across speech styles and speaker groups, but that male speakers were more creaky and female speakers were more whispery, and that working-class speakers were particularly whispery in conversational speech. Beck & Schaeffler [5] also found that whispery voice was prevalent across speaker groups in their study on 76 adolescent speakers in Dumfries, Aberdeen, and Inverness. Other aspects of laryngeal voice quality varied by gender and region, with male speakers being rated higher on scales for creaky voice and harsh voice, and Aberdeen speakers being rated higher in terms of larynx tension than speakers from other areas. The use of VPA in research on Scottish accents suggests that the presence, magnitude, and combination of different laryngeal settings are all relevant for the characterising of sociophonetic variation in Scottish voices.

In this research, I consider voice quality in speakers of Scots originating from Glasgow, Lothian (Edinburgh and surrounding areas), and Orkney and Shetland, referred to here as 'Insular'. This allows comparison with previous research that focuses on the Central Belt varieties of Scots spoken in Scotland's urban centres, while expanding the scope of research to include Insular Scots speakers. Future research should consider working with the wider range of Scots dialect areas available in SCOSYA, which spans a wider range of areas including Dumfries & Galloway, the Western Isles, and the North East.

1.2. Acoustic measurement

Auditory-perceptual schemes for analysing voice quality have been criticised on the grounds of low inter-rater reliability, attributed to the difficulty of isolating individual components of voice quality from the holistic impression of a voice, and rating the continuous perceptual properties on ordinal scales [14]. Acoustic methods side-step these issues while allowing the scale of analysis to be increased, but capturing different components that constitute voice quality acoustically in spontaneous speech collected outside of a laboratory environment can produce challenges. Szakay & Torgersen [15] analysed voice quality in London adolescents using a single measure, H1-H2, but reached different conclusions about how creaky voice patterned in their sample upon revisiting the data using an automated approach to identifying creak [16]. This approach, which involves identifying a speaker-specific f0 antimode that functions as a threshold for separating creaky and non-creaky speech [7, 17], shows promise as a method of producing coarse-grained estimates of the amount of creak that speakers use. In [6], I used this method to consider how creak varies according to social and linguistic factors in 95 speakers of Glasgow, Lothian, and Insular Scots and found that speakers of Insular Scots were particularly creaky. Examining creak in isolation does not approach the level of descriptive detail afforded by VPA: I therefore propose that conducting further acoustic analysis on automatically coded non-creaky voice may be necessary to approximate the insights afforded by VPA.

Drawing on Kreiman et al.'s psychoacoustic model [9, 10], Garellek [11] argues that combining multiple measures of voice quality that relate to the inharmonic noise source and spectral tilt captures the perceptual quality of the voice more adequately than a single measure in isolation. In the present study, I therefore build on [6] and consider how voice quality in non-creaky speech varies according to age, gender, and area in this same sample of 95 Scots speakers, where creak has been removed using the automated approach described in [6], following [7,16,17]. In doing so, I propose that spectral tilt and inharmonic noise measures can then be interpreted more clearly, with spectral tilt being interpretable in terms of tense-lax glottal constriction, and measures of noise relating more likely to turbulent airflow rather than f0 irregularity. I conduct this analysis with the following measures: H1*−H2*, H2*−H4*, H4*−2kHz*, and CPP. For brevity, only H1*−H2* and CPP are presented here. Results for higher-frequency spectral measures (H2*−H4* and H4*−2kHz*) are given in [18].

H1*−H2* is a measure of the degree of glottal constriction. Higher airflow, a longer open phase, and less contact between the vocal folds lead to H1 being more dominant in the low-frequency portion of the spectrum; this in turn leads to less constricted glottal states displaying higher values for H1*−H2* [19]. Meanwhile, CPP is a measure that involves comparison of the amount of aperiodic noise and harmonic energy in the cepstral domain, the log-frequency spectrum [20]. Higher values indicate more harmonic energy in the signal and a more modal-like quality, while lower values indicate more aperiodic noise in the signal, and a less modal quality.

Unlike in VPA, where voice quality is approached as a long-term setting of speech, acoustic measurement is often implemented in vowels alone. Much of the research validating the use of these measures works only with steady-state vowels in isolation due to practical constraints [9]. Implementing these measures in spontaneous speech rather than isolated vowels therefore brings with it considerations and possibilities. Firstly, the present study expands measurement to include other sonorants, which are likely to affect acoustic measurement, and so it is necessary to control for the duration and linguistic content of a stretch. Nasals, for example, display lower CPP than vowels [21]. Furthermore, the linguistic context may also affect phonation and thus the measurements taken: Glottal stops [22], pre-glottalisation [23], phrase finality [24], and phrase-initial vowel onsets [25, 26] may all result in the presence of creak or other cues to glottalization, in the form of decreased values for spectral tilt measures. Furthermore, in Scottish English, pre-aspiration can occur before word-final voiceless fricatives, leading to increased spectral tilt and inharmonic noise in the spectrum [27], which may also be expected in contexts following aspirated voiceless plosives. Controlling for these linguistic effects, however, provides the opportunity to consider how short-term variation in phonation due to segmental content and context compounds to produce the overall voice quality.

1.3. Predictions

I consider variation according to both social and linguistic factors in this data but focus here on variation according to social factors: age (older approx. 65+, younger approx. 18-25), gender (male, female), and area (Glasgow, Lothian, Insular).

1.3.1. Age

Previous work on age in Scottish voice quality has found only minimal differences [3,4]. However, the older (65+) age group considered here are older than the groups considered by [3,4]. Their voices may be affected by ageing processes known as presbyphonia, which may affect voice quality [28]. Laryngoscopy shows 'bowing' of the vocal folds in older speakers and increased prevalence of a posterior glottal gap [29], which results in a breathier voice, manifesting acoustically in the form of increased aperiodic noise in the spectrum [30]. Ageing effects may interact with speaker sex and variety [31, 32]. Looking at British English varieties, Gittelson et al. [32] found that male and female speakers pattern differently with age, with HNR35 and H1*–H2* lowering with age for male speakers but increasing for female speakers. I therefore expect to find age differences, potentially in the form of decreased CPP and/or increased spectral tilt in older speakers but recognise this may not be consistent across all speakers.

1.3.2. Gender

In Glasgow, Stuart-Smith [3] found that male speakers were more creaky and female speakers were more whispery, while Beck, considering mostly south-east Scottish accents [2] and Beck & Schaeffler [5], who looked at adolescents from three areas, found that male speakers were more creaky and more harsh. The present study builds on [6], who found no gender difference in terms of creaky voice in this data. I therefore expect I may find gender differences between male and female speakers, which could take the form of increased whispery voice in female speakers. If female speakers are more whispery than male speakers in their non-creaky voice, this can be expected to manifest in terms of lower CPP, and steeper spectral slope. However, considering that in [6] I found no evidence of differences in terms of creak, contrary to Stuart-Smith [3], Beck [2], and Beck & Schaeffler [5], the findings of this research may continue to depart from previous research on Scottish accents due to a difference in methodology.

1.3.3. Area

VPA research suggests prevalence of whispery voice across Scottish accents [2-5], with harsh voice in male speakers from Edinburgh [4] and tense whispery voice in Glasgow [3]. Acoustically, harsh voice would manifest with particularly low CPP. The acoustic manifestation of tense, whispery voice is unclear: We can speculate that it may involve a combination of characteristics of tense and whispery voice, perhaps involving low spectral tilt alongside low CPP.

2. Data and Methods

I investigated variation in voice quality according to social and linguistic factors in 95 Glasgow, Lothian, and Insular Scots speakers taken from SCOSYA [8]. The analysis presented here builds on the analysis presented in [6], where f0-based coding of creak was used to separate out creaky voice from non-creaky voice and estimate how use of creak varies according to social and linguistic factors. The present analysis then considers all *non-creaky* voice identified in [6]. Controlling for linguistic effects, here I investigate how H1*–H2* and CPP vary according to age, gender, and area.

2.1. Data

Speech data for 95 speakers were taken from the SCOSYA [6], a corpus originally collected to investigate syntactic variation by Scots variety. For each speaker, 180 seconds of conversational speech was considered. The participants are speakers whose spoken varieties occupy the linguistic continuum between Scottish Standard English and Scots [33, p. 4]. Table 1 summarises the sample, which included 19 speakers of Insular Scots (Orkney and Shetland), 29 speakers from Lothian (Edinburgh and surrounding areas), 47 speakers from Glasgow (incl. Greater Glasgow), of which 46 were male, 49 female, 46 older and 49 younger.

Table 1: *summary of sample by age, gender and dialect area.*

	Older (approx. 65+)		Younger (approx. 18-25)		
	Female	Male	Female	Male	Total
Glasgow	10	10	13	15	48
Insular	9	3	3	4	19
Lothian	7	7	7	7	28
Total	26	20	23	26	95

Speakers were sampled through an auditory assessment of background noise and recording quality. The varying number of speakers included from each area reflects the amount of total and usable recordings from each area. Time-aligned transcriptions provided by SCOSYA were converted to TextGrids using [34], then force aligned using the Montreal Forced Aligner v.1.1 [35,36]. The output of the forced aligner was hand-checked and gross alignment errors were excluded or corrected. The forced alignment was then used to identify stretches of sonorants and to estimate linguistic effects such as glottalization.

2.2. Methods

2.2.1. Separation of creak and unit of analysis

The analysis presented in [6] was conducted on stretches of sonorant sounds, termed 'voiced stretches'. Full details of the procedure for identifying creak are given in [6, 18], but in summary, the procedure involves processing sound files using REAPER [37] to identify Glottal Closures Instants (GCIs), which are used to calculate f0. An f0 distribution is then generated for each speaker using an automatic procedure developed by Dallaston & Docherty [7]. This involves identifying peaks in the distribution corresponding to probable modes of creaky and non-creaky f0, and along with a speaker-specific antimode in the trough between them that is then used to separate creaky and non-creaky voicing.

© 2024 Joe Pearce. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

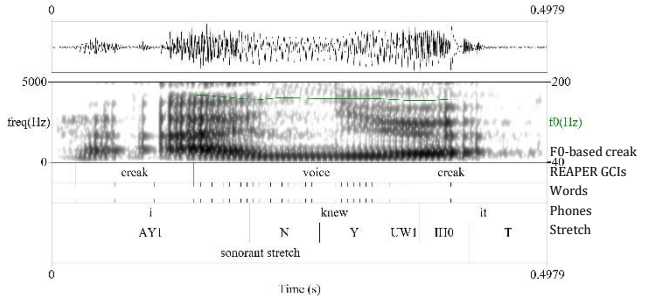


Figure 1: *output of automated identification of creak in sonorant stretches in the speech of a younger female speaker from Glasgow.*

Use of this procedure resulted in creaky and non-creaky stretches being separated, as shown in Figure 1. The bottom tier shows the stretch before identification of creak, while the top tier shows the voiced stretches after identification of creak. The non-creaky stretches identified by this procedure are those that are analysed here.

In the analysis presented in [6], antimodes could not be identified for three speakers. These three speakers are included in the analysis presented here, and all voiced stretches treated as non-creaky. 26,922 non-creaky voiced stretches were analysed, totalling 57 minutes and representing 85% of the total number of stretches considered in [6].

2.2.2. Acoustic analysis

H1*−H2*, H2*−H4*, H4*−2kHz*, and CPP were chosen for consideration following a linked auditory-perceptual and acoustic study of a smaller sample of 24 participants, described in [18], which indicated these measures captured perceptually-relevant variation in voice quality in this corpus. Acoustic analysis was conducted using VoiceSauce [38]. Measurements of H1*−H2*, H2*−H4*, H4*−2kHz*, and CPP were taken using a 16 kHz sampling rate and averaged within each stretch. Harmonic measurements were corrected for formant influences using the formant values and bandwidths estimated using Snack.

2.2.3. Estimation of linguistic effects

Forced alignment of the data all allowed for a number of linguistic effects on phonation to be controlled for without needing to be coded by hand. For example, glottalisation contexts were coded in cases where pre-glottalisation or glottal replacement of /t/ were possible, as outlined by [39, p. 184]. To emphasise that these are possible rather than hand-checked cases of glottal stops, cases where /t/ is likely to be produced as a glottal stop are termed 'glottalable' contexts (e.g. *bet #, betting, button*). Aspiration contexts were coded in stretches that followed voiceless aspirated plosives (e.g. *pine,* but not *spine)*, and voiceless glottal fricatives (e.g. *have*). Similarly, pre-aspiration contexts were coded in stretches containing vowels that precede non-word-initial voiceless fricatives (e.g. *grass*). Phrase position was estimated based on whether the forced aligner detected a pause preceding or following stretch, and paired with considering phonemic content of the stretch, which allowed estimation of context containing phrase-initial vowels. The phonemic content of the stretch also allowed consideration of whether the stretch contained nasals, /l,w,j/ or rhotics.

Table 2: *results of the mixed-effects model predicting H1*−H2* as a function of social and linguistic factors.*

	Coefficient	SE	t	
Intercept	0.362	0.078	4.660	***
Duration	−0.021	0.009	−2.241	*
Glottalable	−0.052	0.018	−2.878	**
Pre-glottalisation	−0.066	0.039	−1.695	
Asp. /p,t,k/	0.018	0.015	1.216	
Asp. /h/	0.105	0.030	3.516	***
Pre-aspiration	0.079	0.024	3.343	***
Both init. & non	0.075	0.050	1.514	
Initial	0.065	0.029	2.256	*
No vowel	−0.280	0.047	−5.935	***
Nasal	−0.147	0.015	−9.572	***
/l,w,j/	0.032	0.014	2.256	*
Female	−0.354	0.084	−4.209	***
Younger	−0.234	0.084	−2.776	**
			Observations = 27,563	
			Log. Likelihood = −36,404.920	

2.1. Statistical modelling

The effect of social and linguistic factors on H1*−H2* and CPP was modelled using mixed-effects linear regression using `lmer` in lme4 (v1.1-28) [40] and in R v4.1.2 [41], with random intercepts for Words (string of orthographic words in the stretch) and Participant. Each model was stepped up manually starting from a minimal model containing only the random intercept for Participant, with factors added sequentially, and random slopes considered for Duration and Speech Rate. Models were compared in a series of log-likelihood ratio tests, where variables that did not significantly improve the fit of the model were not included. The full list of factors considered for inclusion is given in [18].

Fixed effects included in the final models included the Duration of a stretch (Log, scaled), Glottalisation (*None*, Pre-glottalisation, Glottalable), Aspiration (*None*, Aspiration from preceding /p,t,k/, Aspiration from preceding /h/), Preaspiration (*None*, Pre-aspiration), Phrase-Final (*Non-final*, Final), Contains Vowel

(*Non-initial vowel*, No vowel, Initial vowel, Both initial and non-initial) Contains Nasal (*No nasal*, Nasal), Contains Rhotic (*No rhotic*, Rhotic), Contains /l,w,j/ (*No /l, w,j/*, /l,w,j/), Gender (*Male*, female) and Age (*Older*, younger).

The final models were as follows:

1. H1*−H2* ~ Duration + Glottalisation + Aspiration + Preaspiration + Vowel + Nasal + /l,w,j/ + Gender + Age + (1 + Duration || Participant) + (1 | Words)
2. CPP ~ Duration + Aspiration + Final + Vowel + Nasal + Rhotic + /l,w,j/ + Age + (1 + Duration || Participant) + (1 + Duration || Words)

No interactions significantly improved the fit of the model and were therefore not included. This could be related to limitations in the sample described in Table 1, which is relatively even in Glasgow and Lothian but skewed towards older female speakers for Insular Scots speakers.

3. Results

For brevity, only results for H1*−H2* and CPP are given here, and focus is only on social factors. [18] includes additional results for H2*−H4* and H4*−2kHz* alongside more information about these models. Table 2 shows the results of the model predicting H1*−H2* as a function social and linguistic factors.

As shown in Figure 2, H1*−H2* was lower for Younger speakers than Older speakers, and lower for Female speakers than Male speakers. This suggests that in non-creaky voicing, Female speakers and Younger speakers are using more constricted glottal configuration.

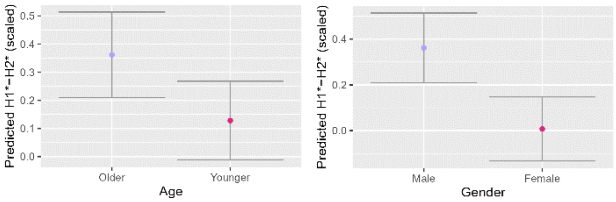


Figure 2: *effect of Age and Gender of H1*−H2**.

Table 3: *results of the mixed-effects linear regression model predicting CPP as a function of social and linguistic factors.*

	Coefficient	SE	t	
Intercept	0.109	0.061	1.784	
Duration	0.440	0.013	32.680	***
Asp. /p,t,k/	−0.079	0.014	−15.744	***
Asp. /h/	0.039	0.033	1.189	
Phrase-final	−0.202	0.012	−17.121	***
Both init. & non	−0.230	0.047	−4.876	***
Initial	−0.287	0.027	−10.687	***
No vowel	−0.496	0.046	−10.817	***
Nasal	−0.393	0.016	−24.391	***
Rhotic	−0.104	0.017	−6.142	***
/l,w,j/	−0.140	0.015	−9.270	*
Younger	0.273	0.083	3.304	**
			Observations = 27,563	
			Log Likelihood = −33,151.110	

Table 3 shows the effects of social and linguistic factors on CPP. As shown in Figure 3, Younger speakers show significantly higher CPP, suggesting use of a more modal-like quality.

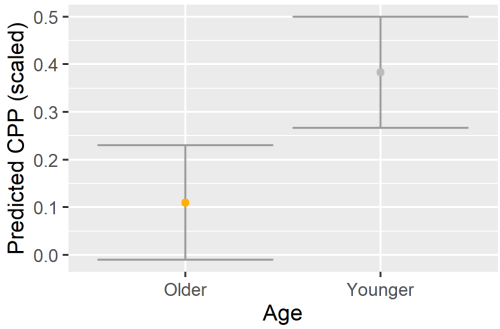


Figure 3: *effect of Age on CPP.*

4. Discussion

How do acoustic measures of voice quality pattern according to age, gender, and area in non-creaky voice in Scottish accents? I found that in non-creaky voicing, female speakers showed evidence of tenser voicing than male speakers, younger speakers display a more modal-like quality than older speakers, and no significant difference by area.

4.1. Age

I expected to find age differences and posited that this would likely be in the form of decrease CPP and/or increased spectral tilt among older speakers, suggesting increased breathiness. Although I predicted ageing effects might not be consistent across age and dialect area, I was unable to consider interactions between these factors.

The findings of the present study show evidence of increased breathiness with age, as captured here in lower spectral tilt and higher CPP among younger speakers. While Gittelson et al. [32] found gender-specific ageing effects on the voice in British English accents more generally, the findings here suggest similar patterning between male and female speakers in Scottish accents. As Hejná et al. [42] emphasise, differentiating between biological and social reasons for age effects is not always straightforward. It is possible that higher CPP and lower spectral tilt among younger speakers reflects a shift away from harsh and whispery voice quality found by [2-5], and equally possible that these differences between older and younger speakers reflect the progression of physiological ageing processes. However, it is

© 2024 Joe Pearce. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

worth noting that Beck & Schaeffler [5] found whispery and harsh voice among Scottish adolescents in the most recent auditory-perceptual study of Scottish voice quality, suggesting this difference is more likely due to the physiological effects of ageing.

4.2. Gender

Considering that Stuart-Smith [3] found that female speakers were more whispery in Glasgow, I expected that I might find increase whispery voice in female speakers in the form of lower CPP and steeper spectral slope. The finding that female speakers in this corpus show greater evidence for tenser or more constricted voice quality is therefore somewhat surprising. In [6], I found no gender difference in creak use in this sample. Taken together, the present research emphasises the need to differentiate the contributions of glottal tension, whisperiness and the use of creak to overall voice quality.

This finding gives counter evidence to the claim that increased breathiness in female speakers is universal. [43] discuss how H1*−H2* is often greater for female than male speakers but that the magnitude of this difference varies between languages, suggesting that variation in breathiness is language-specific. Finding that H1*−H2* is lower in female speakers than male speakers in Scottish accents suggests that gender variation in glottal tension is not only language-specific, but accent-specific, considering that in British English speakers more generally, [32] found H1*−H2* was lower for male speakers than female speakers.

This in turn suggests that if there is a physiological influence that favours increased breathiness in female speakers, as suggested by [44], this is not deterministic. The full picture of voice quality requires taking the relationship between multiple variables into account. Future research should consider using alternative modelling techniques in order to take these complex relationships into account, as modelling them separately here did not allow consideration of how the relationship between two acoustic measures varied according to independent variables.

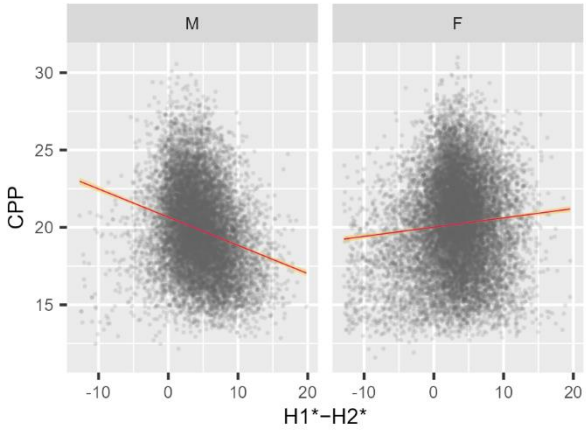


Figure 4: *exploratory scatterplot of the relationship between H1*−H2* and CPP by gender. The pink line shows linear regression fit through the data, with shaded yellow area showing 95% CIs.*

A tentative exploratory analysis, shown in Figure 4, suggests the relationship between H1*−H2* and CPP may differ by gender. Among male speakers, there appears to be a weak negative relationship, with CPP decreasing as H1*−H2* increases. This pattern suggests that noisier stretches may tend to be produced with an open glottal configuration, such that they would be consistent with breathy voice, while less noisy stretches are produced with a more closed glottal configuration, consistent with modal to tense voice. On the other hand, female speakers show a weak positive relationship between CPP and H1*−H2*, with CPP increasing as H1*−H2* increases. Noisy tokens for female speakers appear to show a more closed glottal configuration, which would be more consistent with whispery or tense whispery voice. This suggestion is only exploratory and should not be taken as an inferential statistical analysis. Future research should consider the relationships between different acoustic variables more closely, as this analysis is very preliminary.

4.3. Lack of a difference for area

Previous VPA research on Scottish accents suggests that a degree of whispery voice in common across accents [2-5]. Research that considers different Scottish accents independently suggests that harsh voice is prevalent in male speakers in Edinburgh [4], and tense whispery voice is characteristic of speakers in Glasgow [3]. I expected I might find evidence of these qualities in the form of voice that would manifest with particularly low CPP for harsh voice, and combinations of spectral characteristics of tense and whispery voice for tense whispery voice.

Contrary to these expectations, I found no acoustic differences in voice quality between Glasgow, Edinburgh, and Insular varieties in non-creaky voicing. This finding is, however, broadly consistent with the only study to directly compare different Scottish accents by Beck & Schaeffler [5], who found minimal differences by area between Aberdeen, Inverness, and Dumfries [5].

Though it may be that changes over time mean that there is no longer a difference in voice quality between Glasgow and Edinburgh speakers, an alternative explanation may be that an acoustic, corpus-based approach is more suited to identifying some types of variation than others. Harsh voice, though easily identifiable auditorily, might be more difficult to identify through purely acoustic methods, and is likely to occur intermittently, meaning that instances of harshness might be present in the data but not clearly visible. Forms of harsh voice that involve aryepiglottic trilling may also involve low pitch [45, p. 74], which would mean that they would likely have been identified as creak by the automated procedure in [6]. Another explanation may be that differences in voice quality between Glasgow and Edinburgh might come in how different measures pattern together, rather than in single measures alone.

4.4. Limitations and future directions

A limitation of the present study comes in the use of the f0-based approach. As summarised by [46], not all types of creak are characterised by low f0, with aperiodicity and multiple pulsing

also characterising some creak. These would not necessarily be identified as having a low f0 within a speaker's range, and in turn would have been analysed here as non-creaky voice. Furthermore, it is worth noting a limitation of this approach specifically for considering variation according to social factors. White et al. [47] found that this method performed worse for male speakers than for female speakers, while preliminary work by Pearce [18] found that this method was less reliable for older speakers than younger speakers. It is therefore possible that some findings here that appear to relate to age and gender may in fact reflect differences in how much creak is reliably identified by speaker group. However, the present study has reinforced the need to consider creak as part of a wider multidimensional voice. Future research may therefore consider a similar approach that conducts subsequent acoustic analysis on creaky and non-creaky speech after hand-coding of creaky voice, or which uses an improved method of automatically detecting creaky voice. For example, White et al. [47] found that combining the low-f0 approach with Kane et al.'s resonator-based creaky voice detection algorithm [48] brought automatic detection of creak closer to hand-coded detection.

5. Conclusion

This acoustic analysis found that non-creaky laryngeal voice quality in Scottish accents varies according to age, with younger speakers producing a more modal quality, and gender, with female speakers producing a tenser quality.

The patterns uncovered here demonstrate the potential use of a multi-pronged, multi-measure approach in sociophonetic studies of voice quality. If creaky voice had been considered alongside non-creaky voice, I might have connected lower H1*−H2* among younger speakers and female speakers to creaky voice. By comparison, the approach taken here reveals similarity between male and female speakers in the amount of creak used and gives greater insight into the voices of speakers when they are not using creak - which for these speakers, is most of the time. Considering the separate contributions of creak, inharmonic noise, and glottal constriction to voice quality in unison reveals that Scottish female speakers use a more tense, but *not* more creaky, quality than male speakers. Further, exploratory post-hoc analysis indicated that noisy voice qualities produced by male and female speakers may be produced with separate laryngeal configurations, with female speakers potentially favouring whispery voice and male speakers potentially favouring breathy voice. Though the approach taken here has limitations, particularly in terms of its ability to differentiate harsh voice from other constricted laryngeal states, this research presents a step towards quantifying how individual components of voice quality contribute to overall voice quality in an acoustic, corpus-based study of spontaneous speech. Future research may gain greater insight into laryngeal configurations underlying sociophonetic variation in voice quality by focusing on the relationships between different acoustic measures.

6. Acknowledgments

This research was funded by the ESRC through SGSSS (Project reference: 2178789). I would like to thank Jane Stuart-Smith, Clara Cohen, and Felix Schaeffler for their supervisory input on this project.

7. References

1. Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*, 14, 139-155.
2. Beck, J. M. (1988). *Organic variation and voice quality*. University of Edinburgh dissertation.
3. Stuart-Smith, J. (1999). Glasgow: Accent and voice quality. *Urban voices: Accent studies in the British Isles,* ed. P. Foulkes & G. Docherty, 203-222. Springer, J. (2019).
4. Esling, J. (1978). The identification of features of voice quality in social groups. *Journal of the International Phonetic Association, 8*(1-2), 18-23.
5. Beck, J. M., & Schaeffler, F. (2015). Voice quality variation in Scottish adolescents: gender versus geography. *Proceedings of the 18th ICPhS, Glasgow*.
6. Pearce, J. (2023). Creaky voice in three Scots varieties: Using f0-based identification to consider social and linguistic factors. In: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, 816-820. Guarant International.
7. Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. In *Proceedings of the 19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association Inc* (pp. 532-536).
8. Smith, Jennifer; Adger, David; Aitken, Brian; Heycock, Caroline; Jamieson, E & Thoms, Gary. (2019). The Scots Syntax Atlas. University of Glasgow. https://scotssyntaxatlas.ac.uk
9. Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1), e009. https://doi.org/10.3989/loquens.2014.009
10. Kreiman, J., Lee, Y., Garellek, M., Samlan, R., & Gerratt, B. R. (2021). Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America*, *149*(1), 457-465.
11. Garellek, M. (2019). The phonetics of voice. In W. F. Katz & P. F. Assmann (eds.) *The Routledge Handbook of Phonetics,* 75–106. Oxford: Routledge.
12. Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press.
13. Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.
14. Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*, *122*(4), 2354-2364.
15. Szakay, A., & Torgersen, E. (2015). An acoustic analysis of voice quality in London English: The effect of gender, ethnicity and f0. In *Proceedings of the 18th ICPhS, Glasgow*.
16. Szakay, A., & Torgersen, E. (2019). A re-analysis of F0 in ethnic varieties of London English using REAPER. I *Proceedings of the 19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association Inc* (pp. 1675-1678.).

© 2024 Joe Pearce. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

17. Dorreen, Keiran. (2017). *Fundamental frequency distributions of bilingual speakers in forensic speaker comparison*. University of Canterbury MA thesis.
18. Pearce, J. (2023). Social meaning in Scottish voices: A sociophonetic investigation of voice quality combining auditory-perceptual, acoustic, and qualitative approaches. PhD thesis. University of Glasgow.
19. Zhang, Z. (2016). Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *The Journal of the Acoustical Society of America*, *139*(4), 1493-1507.
20. Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4), 769-778.
21. Madill, C., Nguyen, D. D., Yick-Ning Cham, K., Novakovic, D., & McCabe, P. (2019). The impact of nasalance on cepstral peak prominence and harmonics-to-noise ratio. *The Laryngoscope*, *129*(8), E299-E304.
22. Stuart-Smith, J. (1999). Glottals past and present: A study of T-glottalling in Glaswegian. *Leeds Studies in English*, 181-204.
23. Gordeeva, Olga B., & James M. Scobbie. (2013). A phonetically versatile contrast: Pulmonic and glottalic voicelessness in Scottish English obstruents and voice quality. *Journal of the International Phonetic Association*, *43*(3), 249-271.
24. Henton, C. & Bladon, A. (1988). Creak as a sociophonetic marker. In L M Hyman & C N Li (eds.), *Language, speech and mind: studies in honour of Victoria A. Fromkin*, pp. 3–29.
25. Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. Journal of phonetics, 24(4), 423-444.
26. Pierrehumbert, J. & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Doherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture Segment Prosody*, 90–117. Cambridge: Cambridge University Press.
27. Gordeeva, O. B. & Scobbie, J. M. (2010). Preaspiration as a correlate of word-final voice in Scottish English fricatives. English. In Susanne Fuchs, Martine Toda, Marzena Żygis & ProQuest (Firm) (eds.), *Turbulent sounds: an interdisciplinary guide,* vol. 21;21.; 167–207. Berlin;New York, N.Y; Mouton de Gruyter.
28. Kendall, K. (2007). Presbyphonia: a review. *Current opinion in otolaryngology & head and neck surgery*, *15*(3), 137-140.
29. Martins, R. H. G., Benito Pessin, A. B., Nassib, D. J., Branco, A., Rodrigues, S. A., & Matheus, S. M. M. (2015). Aging voice and the laryngeal muscle atrophy. *The Laryngoscope*, *125*(11), 2518-2521.
30. Rojas, S., Kefalianos, E., & Vogel, A. (2020). How does our voice change as we age? A systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research*, *63*(2), 533-551.
31. Lee, S. J., Cho, Y., Song, J. Y., Lee, D., Kim, Y., & Kim, H. (2016). Aging effect on Korean female voice: Acoustic and perceptual examinations of breathiness. *Folia Phoniatrica et Logopaedica*, *67*(6), 300-307.
32. Gittelson, B., Leemann, A., & Tomaschek, F. (2021). Using crowd-sourced speech data to study socially constrained variation in nonmodal phonation. *Frontiers in Artificial Intelligence*, *3*, 565682.
33. Stuart-Smith, J. (2008). *Scottish English: Phonology*. In Kortman; Upton (eds.). Varieties of English: The British Isles. Mouton de Gruyter, New York.
34. Fromont, R. (2022). *trs2grid.jar*. [Online]. Available: https://sourceforge.net/projects/labbcat/
35. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, & M. Sonderegger. (2017). "Montreal Forced Aligner: trainable text-speech alignment using Kaldi," in *Proceedings of the 18th Conference of the ISCA*, 2017.
36. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, & M. Sonderegger. (2017). "Montreal Forced Aligner [computer program]. v.1.1.0," 2017. [Online]. Available: http://montrealcorpustools.github.io/Montreal-Forced-Aligner/
37. Talkin, D. (2015). https://github.com/google/REAPER
38. Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis, *Proceedings of the ICPhS XVII*, 1846-1849.
39. Gimson, A. C & Cruttenden, A. (2014). Gimson's Pronunciation of English. 8th edn. London: Routledge.
40. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1). 1–48.
41. R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria
42. Hejná, M., & Jespersen, A. (2022). Ageing well: Social but also biological reasons for age-grading. *Language & Linguistics Compass*, e12450. https://doi.org/10.1111/lnc3.12450
43. Hejná, M., P., Tylečcová, L. & Bořil, T. (2021). Normophonic breathiness in Czech and Danish: are females breathier than males? *Journal of Voice 35(*3). 498.e1–498.e22
44. Titze, I. R. (1989). A four-parameter model of the glottis and vocal fold contact area. *Speech communication*, *8*(3), 191-201.
45. Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). *Voice quality: The laryngeal articulator model*. Cambridge University Press.
46. Keating, P., Garellek, G. & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In *Proceedings of the 18th ICPhS, Glasgow*.
47. White, H., Penney, J. Gibson, A., Szakay, A. & Cox, F. (2022). Evaluating automatic creaky voice detection methods. *Journal of the Acoustical Society of America* 152(3). 1476–1486.
48. Kane, J., Drugman, T. & Gobl, C. (2013). Improved automatic detection of creak. *Computer Speech & Language,* 27(4). 1028–1047.

DOI: 10.2478/9788366675513-005

“Vocal interactions”: the use of voice in spoken word duos

Rebecka Dürr[1]

[1]Institute for German Language and Literature, University of Hamburg, Germany

rebecka.duerr@uni-hamburg.de

ABSTRACT

This article provides insights into the use of the voice in spoken word duos, with a focus on the German language. It considers vocal interactions in spoken word poetry in relation to both the spoken word movement and more performative approaches to poetry. This includes various multicultural, literal, and musical influences, and also expands our understanding of poetry against the backdrop of literary competitions. In order to introduce, explain, and discuss the different ways that vocal interaction is used as a creative means and in relation to the content of slam texts, the article takes a speech communication and literary studies perspective. Artists like *Team Scheller* and *Team Zum goldenen Schmied* use rhythmic patterns, beatboxing, body percussion, and different vocal modes that oscillate between speaking and singing to structure their performances vocally, rhythmically, and physically. This article presents some of the different articulatory, prosodic, and creative means they employ to organize, shape, and perform slam texts and explains how two artists interact with each other to manifest their various intentions and to create situations and sonic impressions.

Key words: spoken word duos, slam poetry, speech communication, literary studies, beatboxing, speech expression, German language

1. Introduction

Most poetry slam contributions are performed by individuals, whereas only a few duos take the stage. Competing in teams, however, not only doubles the number of performers on stage but also multiplies the possible uses of verbal and vocal creative means, including mutual interaction. Slam poets and duos might improvise or recite their poems spontaneously, but in most cases, they spend plenty of time composing, remembering, choreographing, and practicing both texts and performances. Due to the strong entanglement between texts and their verbalization, it is constructive to consider the use of voice in such duos from the perspective of speech communication: as a medium of both oral expression and social interaction simultaneously. This understanding foregrounds the relationship between linguistic formulations, the communicative situation, and verbal expression, rather than the material or phenomenological aspects of voice.

This article will provide insights into the use of the voice in spoken word duos by looking at the German-speaking duos *Team Scheller* and *Team Zum goldenen Schmied*. I will discuss the different ways in which they interact vocally in order to show how they use singing, humming, speaking, and beatboxing to perform slam poetry. This mode of performance imitates certain musical styles and uses rhythmic patterns and speech melodies to generate tension between form, function, and content. In this context form might be a particular rhythmic pattern, while function refers to the use of a creative device to, for example, structure, accompany, or characterize words, phrases, or whole sections of text. When combined, form and function condition the performance of the content. In order to outline this interrelation, this article will take a speech communication perspective to describe the different levels of form and function while considering the content and interpretive approaches from the perspective of literary studies.

At least two aspects are of particular interest: first, the use of different articulatory, prosodic, and creative means to organize, shape, and perform the slam text; and second, the different ways in which the two artists on stage interact with each other to manifest their various intentions and create situations and sonic impressions.

2. Spoken word duos: expanding creative means

Considering vocal interaction in spoken word duos means looking at both the spoken word movement as well as more experimental and performative approaches to poetry. Spoken word encompasses a range of multicultural, literary, and musical influences, which has led to the emergence of a wide variety of formats and helped spoken word poetry to flourish far beyond the borders of the United States [1]. With poetry slams as events

© 2024 Rebecka Dürr. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

where slam poetry is performed, the oral tradition of spoken word, and hereby of witnessing and recounting cultural, historical, and social conditions is carried forward, while at the same time being adapted to address the modern human condition [2]. Poetry slams have been gaining in popularity in Germany since the early 1990s and have now become perhaps the most famous spoken word format. Their rise in popularity has not only been driven by their traditional inclusion in literary and cultural institutions but has also more recently benefited from the opportunities presented by the streaming and the archival features of online platforms like Vimeo or YouTube [3]. So does this article, since the discussed examples are live recordings of slam performances made available on YouTube. Thus, poetry slam culture "mediates the aesthetic qualities of poetry readings with more popular or everyday cultural aspects as well as the competitive character of hip hop and rap culture" [4]. It is therefore not surprising that these various influences are also reflected in the texts performed – referred to as slam poetry – and in modes of interaction.

As for the performed texts, it is difficult to define common formal qualities, modes of address, or general content as slam poets want their poetry to achieve a "more immediate, personal, and authentic engagement with its audience" [5] while they simultaneously compete in a literary competition. However, creative means such as homophonic word play, rhyme, call-and-response, beatboxing, rhythmic structures, and song-like sequences are frequently used in individual and team performances. Being alone on stage presents few options, but team performances open up more possibilities to structure and organize both the text and the staging. They also allow for a more diverse use of vocal and verbal creative means as the interaction between two voices allows performers to alternate their speech complementarily or simultaneously.

In this context, Julia Lajta-Novak's research on live poetry introduces articulatory as well as composite parameters to describe live poetry performances [6]. While she develops her approach in relation to Theo van Leeuwen's publication *Speech, Music, Sound* [7], German literary scholar Frieder von Ammon in his study on Austrian poet and author Ernst Jandl relies on musical categories such as timbre and pitch [8]. The field of speech communication enables a research-oriented specialization of the works and approaches mentioned above, offering a variety of established concepts, terms, and methodological approaches developed specifically in relation to spoken utterances and accompanying interpersonal behaviour. This allows for a more detailed description of speech sounds and their relation to both the spoken text and the situational context (performance) as well as the possibility to attribute aesthetic effects to individual vocal and articulatory parameters.

In this regard, the speaking voice is considered an essential component of oral communication, and simultaneously serves as the medium of aesthetic expression and social interaction. As such it can be described and analysed precisely by using the concept of speech expression (*Sprechausdruck*) as well as associated parameters, which are expanded upon below. The so-called speech expression comprises a complex of timbral and melodic, temporal, dynamic, and articulatory parameters that can be further subdivided into different prosodic means such as pitch, loudness, and complex phenomena like speech rhythm and accentuation [9]. In spoken utterances they appear individually and/or in combination, helping to structure utterances, highlight certain information, affect meaning [10], etc. This dichotomy of form (pitch etc.) and function (structure etc.) helps to shape the spoken word in artistic contexts. Since the resulting shapes and patterns are based on everyday language use, they can evoke emotions, memories, and associations in listeners.

Another communicative behaviour that appears in slam performances is singing, which, just like speaking, serves to convey information, emotions, or the performer's personal state. Even though speaking and singing rely on the same physiological conditions, they differ with regard to the individual use of the speech organs and correspondingly in terms of formal, technical, and tonal characteristics [11]. Nevertheless, it is not possible to draw distinctions between them on the basis of fundamental characteristics alone, even though a variety of phenomenological descriptions refer to pitch movements and intervals as the main qualities in the transition between speaking and singing [12]. While speaking is usually characterised by sliding, constantly changing articulatory movements with indefinite pitch, singing, due to its use of compositionally defined intervals, demonstrates rather gradual tone sequences and articulation that is dominated by vowels [13]. However, smooth transitions between these two types of phonation are common in practice as different artists tend to use both of them to varying degrees and in a process of constant redefinition [14]. Moreover, articulation, formal features predefined by the musical composition, and both situational and contextual circumstances determine the extent to which the listener perceives the text as being spoken or sung. In this sense, speaking and singing form a continuum on which both types of phonation converge, especially in an artistic context, resulting in forms of expression that oscillate between the two extremes [15]. Thus, in spoken word poetry, even the slightest changes in, e.g., articulation result in the possibility of creating various sound impressions that are reminiscent of both music (soul, hip-hop, etc.) and other more traditional forms of poetic recitation.

The use of the aforementioned modes of voice and the respective articulatory as well as prosodic means becomes more complex in spoken word duos. On the one hand, vocal interaction enables a greater variety of sounds in that two or more voices merge, overlap, or alternate. On the other hand, performing together requires a shared focus, intention, the ability to anticipate one another's actions, and mutual trust to coordinate the temporal structure of both verbal and/or physical performance elements. Consequently, these very elements are subject to agreement between the performers as to how to perform the poem verbally, rhythmically, and physically [16].
The following case studies provide examples of the extent to which vocal interaction is used in slam performances, considering some of their forms and functions, and the impressions they help to create.

3. Spoken word and sounding body

The spoken word duo Dalibor Marković and Dominique Macri, aka *Team Scheller*, employ a variety of non-speech sounds in their slam performances, most prominently beatboxing, body percussion, and melody. When performing their slam text "Atem und Hände" (*Breath and Hands*) [17], for example, both slammers are on stage, standing next to each other, facing the audience. Going against prevailing expectations of the spoken word format, the audience first hears a rhythmic-melodic sequence, which recurs throughout the performance (see Fig. 1). Marković introduces a body percussion pattern, which he creates by tapping on his chest, alternating hands. After establishing the rhythm for a period of one-and-a-half bars, he then additionally starts humming a melody. The hummed melody comprises the interval of a fourth, starting with C3, dropping down a third to A2, then to G2, before going back to C3. All tones sound for three beats, followed by a pause of equal length. The pauses are eliminated for the last three bars, resulting in a continuous sound that simultaneously provides an imperceptible cue. Once Marković returns to C3, thereby completing the whole pattern for the first time, Macri breathes in and starts to speak the poem while her partner finishes the tone's last beat. Marković's hummed melody thus functions as structural means of indicating to his partner when to start speaking.

When contrasted with the content of the poem, the rhythmic-melodic pattern also gains a second function, which is to indicate different temporal layers. To begin with, the first lines of the poem reflect upon different emotional states of memory and loss, culminating in Macri asking, "Erinnerst du dich?" (*Do you remember?*). In what follows, the two performers tell the story of a friendship between two people, which begins with their first encounter and continues through an intimate friendship that ends with the death of one of them due to illness. The dynamic of the interaction between the two articulated "I"s, which is depicted on the textual level and embodied by the performers, is also represented in the structural arrangement of the spoken text. While some experiences are rendered as memories, others are depicted in the moment of speaking. In the case of the former, Macri speaks the text, and Marković accompanies it with the aforementioned rhythmic-melodic pattern. The moments of experience are performed by Macri and Marković together, either in dialogue form, by complementing each other, or speaking in sync.

Thus, the rhythmic-melodic pattern is closely linked to the textual expression of memory and remembering. It does not just initiate the performance but also ends it. In this context, it further symbolizes the idea of "being alone": in the beginning, it is because the articulated "I"s have not yet met and, at the end, due to the death of one of them. Consequently, the different uses of vocal interaction help to underline the temporal dimension of the text and to distinguish sonically between current experience and memory. Since the pattern serves as an opening and closing sequence, it provides the performance with an additional frame. The spatial organization supports this entanglement visually. At the end, Macri leaves the stage, while Marković remains alone in front of the audience. The performers thus performatively duplicate the semantic level – with Marković embodying the surviving friend. Unexpectedly and after finishing the last pattern, Macri, from the off-stage, again asks, "Do you remember?" In doing so, she inverts the initial situation. At the beginning of the performance, she embodies the articulated "I" that remains and remembers, while at the end, her disembodied acousmatic voice, in combination with Marković on stage, refers to the fact that he is now the surviving friend who remembers.

Consequently, the rhythmic-melodic pattern develops both the structural and narrative temporality of the performance. On the one hand, it provides cues in order to facilitate the chronological sequencing of verbal, vocal, and performative elements. Then again, its function as a creative means of musical form relates to memory, since it implicitly refers to the passing of time, both to that of the performance itself and to that of the inner textual moments of friendship. The accompanying pattern thus reminds the audience that the performed sections are remembering times past while simultaneously presenting the memories of the articulated "I"s. Finally, the rhythmic-melodic pattern incorporates the narrative of remembering while at the same time serving as a reminder for the artists as well as the audience.

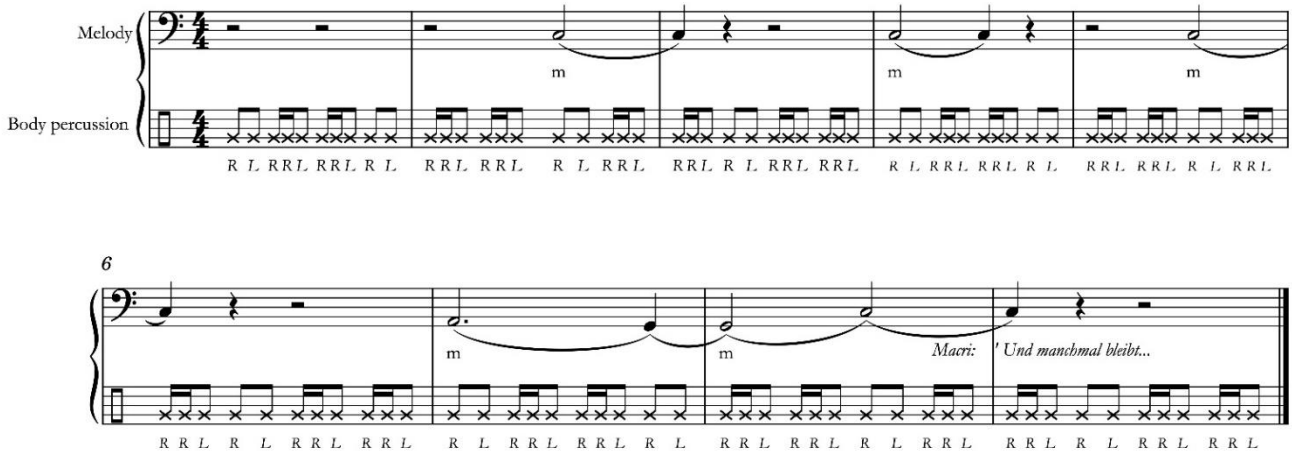


Figure 1: *score of the rhythmic-melodic pattern. Key: R and L indicate the use of either the right or the left hand.*

© 2024 Rebecka Dürr. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

4. Sounding word and soulful music

While in the previous example the spoken mode of voice prevails, Fatima Moumouni and Laurin Buser, aka *Team Zum goldenen Schmied*, combine the practices of beatboxing and call-and-response while alternating between the speaking and singing modes of voice in their performance "Mit Liebe gemacht" (*Made With Love*) [18]. The beginning of the performance is already suggestive of a staged concert. By introducing the team and providing insights into what they write about, Moumouni initially follows the expected course of a slam performance before moving on to give the impression of a concert. The slam poem's content connects the level of the performance and the concert, since every part of the text develops another facet of the topics of writing and slamming. Moumouni starts by pretending that she is testing her microphone and performing a soundcheck. At the same time, Buser again mentions the team name and announces in English that there will be "a little soul music." He furthermore addresses the audience by saying, "I hope you can all embrace it," before turning to his partner. The short cue "Take it" then serves as a signal for Moumouni to start the text. The first phrase "Es sprach einst ein Text zu mir" *(A text once spoke to me*) seems to be more sung than spoken due to its rhythmic composition and melodic structure, including definite pitch. Nevertheless, there is a sentence accent in the phrase, as is usual in speech. The sentence accent marks the stressed parts of a clause, which in this case also functions as the main accent and thus highlights the most important word in the sentence – "text." Moumouni emphasizes this word, which becomes more perceptible as it is louder and has a higher pitch than the surrounding words. Thus, even the first words point to the oscillation between language and music, speaking and singing that characterizes the entire performance. Additionally, the emphasized "text" tells Buser to start making a beatbox pattern, imitating a kick drum {Bm}, snare {^Kh}, and hi-hat sounds {ts}, as well as an implied bass line (see Fig. 2). Reversing the previous example, it is not the musical event that cues when the partner is to start speaking but the emphasized spoken word that marks the start of a musical line.

While the kick drum sound is voiced and therefore sonorous, the snare and hi-hat sounds are produced unvoiced and are therefore percussive. Furthermore, Buser repeats the word "du" three times as part of the beatbox pattern. The context, however, emphasizes its sonic rather than semantic quality, creating a kind of chromatic scale or bass line, as its pitch rises in semitone steps with each repetition – from C4, to C#4, and D4. When he speaks the third "du", Buser combines the bass sound with the kick drum sound, making it possible to either hear it as part of the beatbox element or as part of the bass line. Hence, the combination of voiced and unvoiced, sonorous and percussive elements as well as the characteristic tone sequence creates the illusion that the performers are being accompanied by the drums and bass. The underlying rhythm can be described as six eights. Moumouni additionally visualizes the rhythm with her body, shifting her weight alternately from her left to her right leg, always changing on stressed beats one and four. On these very beats, Buser produces the kick drum and the snare sound, while on unstressed beats two, three, five, and six, he makes the open hi-hat sound. The recurring rhythmic pattern thus invokes particular rhythmic structures referred to as "grooves" that are most prominently used in soul music.

Occasionally, Buser directly comments on or questions the preceding statements made by his partner, incorporating these interactions into his beatbox pattern. This corresponds to the musical scheme of call-and-response, which is found in different musical, cultural, and religious traditions, including soul music. In this sense, Buser's interjections not only expand on the beatbox pattern but also provide a response to Moumouni presenting the main text or, more specifically, the call. Her manner of speaking and singing evolves throughout the performance. While in some parts she creates rhythmic structures reminiscent of rap, in others, she phrases her delivery more musically by using a broader vocal range, definite pitch, and improvised coloratura. In this manner, she articulates the name "Fatima" by lengthening the final vowel <a> and adding a melody. Against the background of the chromatic tone sequence in the beatbox pattern discussed above, Moumouni again uses three halftone steps – from D4, to D#4, to E4 – each used to sing one of the names' three syllables. Thus, the interaction between rhythmic structures, tonal sequences, a certain manner of singing, and the call-and-response pattern together create the impression of soul music. This complex textual, rhythmic, and

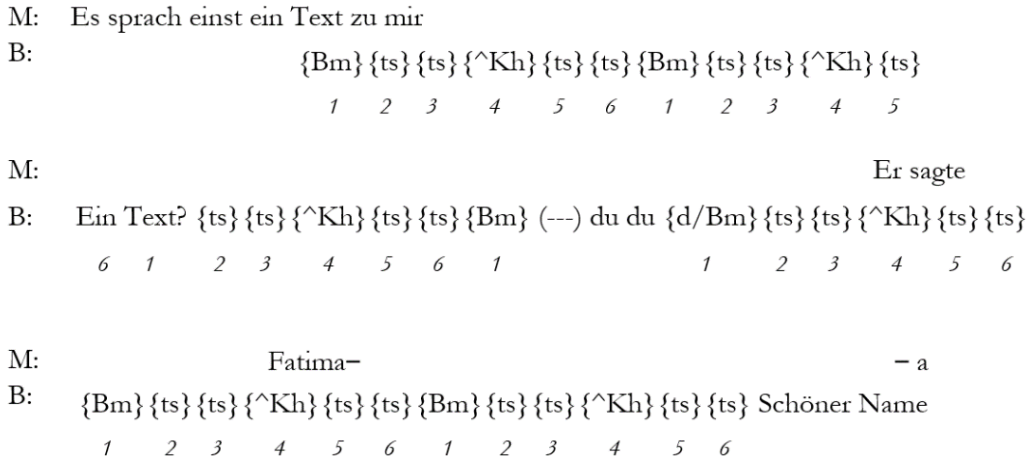


Figure 2: *transcription of the entanglement between beatbox pattern and text. Key: M and B indicate either Moumouni or Buser articulating.*

sonic composition would not be possible without the increased possibilities presented by verbal and vocal interaction.

While in Marković and Macri's performance, vocal interaction mainly serves to create a temporal structure, *Team Zum goldenen Schmied* employ both vocal interactions and different modes of voice to allude to another genre and format. Although it is a slam performance and is introduced as such, the pretence of the soundcheck and the duo's interaction with the audience frame the performance as a concert. An inner performance (concert) is thereby related to an outer performance (slam). Throughout the staging, the team introduces further situations such as TED Talks, sermons, and traditional poem recitations, all of which create the same kind of tension between the inner and outer level. Even though the artists playfully and imperceptibly change the situations and the accompanying vocal styles, they always finish and connect the different sections and facets with one and the same slogan: "Because if you do things with love, you will get love back." By iterating this phrase at the end of each section, they imply that it is not necessarily the type of performance (sermon, recitation, etc.) that is relevant but one's personal relation to, for example, practices such as writing and slamming, as well as the motivation behind them. For a positive, caring, and passionate attitude also communicates with a person sitting opposite or a potential audience.

5. Conclusion

Given the possibility of dividing up the individual passages, the mere act of reciting a text together already expands the sound impressions that a slam poetry performance might evoke. This division might be the most fundamental and simple form of organization, yet it forms a starting point for more complex structures and vocal interactions. The performances discussed in this paper demonstrate an entanglement of vocal, verbal and musical elements. Both duos use articulatory and prosodic means to emphasize and structure oral expression while at the same time referring to different styles of music.

In both examples, the duos constantly change the ways in which they use their interacting voices. This affects parts that are spoken or sung individually, that are supplemented by humming, body percussion or beatboxing, that are spoken in sync, and that show distinct rhythm. In some cases, the constant changes between the speaking and singing modes of voice blur the lines between different performance styles while simultaneously helping to establish those very styles. Unlike solo slam performances, vocal interaction thus creates different temporal, emotional, and situational states without explicitly naming and framing, for example, temporal sequences (*Team Scheller*) or modes of presentation (*Team Zum goldenen Schmied*). Moreover, they help to coordinate the chronological order of verbal, vocal, and rhythmic layers (cues). Both structural functions and aesthetic effects enhance the content by associatively creating sonic references (soul, hip-hop, memory), which either add a layer of meaning to the slam text or differentiate between the individual facets of the content.

The spatial use of the stage complements this approach as well by helping the artists to visualize a range of interpersonal relationships and social interactions. It allows for them to expand on their sound impressions, as a physically disembodied, off-stage voice may evoke other associations than the simultaneous perception of vocal and physical appearance. In the examples discussed here, corporeality, i.e., body movements and gestures, becomes evident in its sensual and material dimension as a performative and musical element while also serving as a tool to support the structural and formal organization of both single sequences and the performance as such. While this article focuses primarily on the use of voice in spoken word duos, the aforementioned range of topics, including corporeality and embodiment, constitutes a fruitful subject for further research on performing in teams.

6. Acknowledgments

This article was written as part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 884177).

Views and opinions expressed here are those of the authors only and do not necessarily reflect those of the European Union (EU) or the European Research Council (ERC). Neither the EU nor the ERC can be held responsible for them.

7. References

1. Westermayr, S. (Ed.). (2010). *Poetry Slam in Deutschland. Theorie und Praxis einer multimedialen Kunstform*. Marburg: Tectum.
2. Smith, M. K. & Kraynak, J. (Eds.). (2009). *Take the Mic. The Art of Performance Poetry, Slam, and the Spoken Word.* Naperville: Sourcebooks.
3. Preckwitz, B. (Ed.). (2005). *Spoken Word & Poetry Slam. Kleine Schriften zur Interaktionsästhetik*. Wien: Passagen.
4. Schweppenhäuser, J. & Pedersen, B. S. (2017). Performing Poetry Slam and Listening Closely to Slam Poetry. *SoundEffects*, 7.1, 64–83.
5. Somers-Willett, S. B. A. (Ed.). (2009). *The Cultural Politics of Slam Poetry. Race, Identity, and the Performance of Popular Verse in America.* Ann Arbor: The University of Michigan Press.
6. Lajta-Novak, J. (Ed.). (2011). *Live Poetry. An Integrated Approach to Poetry in Performance.* Amsterdam, New York: Rodopi.
7. Van Leeuven, T. (Ed.). (1999). *Speech, Music, Sound.* London: Palgrave.
8. Von Ammon, F. (Ed.). (2018). *Fülle des Lauts. Aufführung und Musik in der deutschsprachigen Lyrik seit 1945: Das Werk Ernst Jandls in seinen Kontexten.* Stuttgart: Metzler.
9. Hirschfeld, U. & Neuber, B. (2010). Prosodie im Fremdsprachenunterricht Deutsch – Ein Überblick über Terminologie, Merkmale und Funktionen. *Deutsch als Fremdsprache. Zeitschrift zur Theorie und Praxis des Faches Deutsch als Fremdsprache*, 47.1, 10-17.
10. Hirschfeld, U. & Stock, E. (2013). Sprechwissenschaftliche Phonetik. In Bose, I., Hirschfeld, U., Neuber, B. & E. Stock (eds.), *Einführung in die Sprechwissenschaft. Phonetik, Rhetorik, Sprechkunst* (pp. 27-80). Tübingen: Narr.

© 2024 Rebecka Dürr. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

11. Anders, Y. (2013). Sprecherziehung für Gesangsstudierende. In Bose, I., Hirschfeld, U., Neuber, B. & E. Stock (eds.), *Einführung in die Sprechwissenschaft. Phonetik, Rhetorik, Sprechkunst* (pp. 229-237). Tübingen: Narr.
12. Seidner, W. & Wendler, J. (2005). Sprech- und Singstimme. In Seidner, W., Wendler, J. & U. Eysholdt (eds.), *Lehrbuch der Phoniatrie und Pädaudiologie* (pp. 96-104). Stuttgart: Thieme.
13. Krech, E.-M., Stock, E., Hirschfeld, U. & Anders, L. C. (Eds.). (2009). *Deutsches Aussprachewörterbuch*. Berlin, New York: De Gruyter.
14. Musil, B. (Ed.) (2018*). »Wie ein Begehren«. Musik in der Interpretation von Vokalmusik (nicht nur) des Fin de Siècle*. Bielefeld: Transcript.
15. Merrill, J. (2019). Perzeption und Rezeption des vokalen Ausdrucks im Grenzbereich von Singen und Sprechen. *Sprechen. Zeitschrift für Sprechwissenschaft, Sprechpädagogik, Sprechtherapie, Sprechkunst*, 68, 42-58.
16. Pfänder, S., Herlinghaus, H. & Scheidt, C. E. (2017). Synchronisation in Interaktion. In Breyer, T., Buchholz, M. B., Hamburger, A., Pfänder, S. & E. Schumann (eds.), *Resonanz – Rhythmus – Synchronisierung: Interaktionen in Alltag, Therapie und Kunst* (pp. 65-83). Bielefeld: Transcript.
17. Team Scheller. (2018). *Atem und Hände*. https://www.youtube.com/watch?v=bh5wMtwoVoo (10. August 2023).
18. Team Zum goldenen Schmied. (2016). *Mit Liebe gemacht*. https://www.youtube.com/watch?v=nzGO8N55NJc (10. August 2023).

DOI: 10.2478/9788366675513-006

The 'Rebellious Girl Who Doesn't Try Too Hard': indexicality and independence in the music of Lily Allen

Josephine Iben Kjeldbjerg[1]

[1]Department of English, Aarhus University, Denmark

202105983@post.au.dk

ABSTRACT

The dominance of an American singing style in British music has long been acknowledged and different approaches have been applied to both determine the motivations behind singers' convergence to this style modification but also to determine why some choose to diverge from this norm. Previous research has mainly focused on theories of genre appropriateness and accommodation towards admired groups or different audiences, but these theories have not been sufficient in offering a full understanding when looking at motivations behind the London-born singer-songwriter, Lily Allen's, linguistic behaviour. Six of Allen's songs were analysed in order to provide a quantitative dataset for the claim that Allen portrays Estuary English in her singing style. Moreover, non-linguistic dimensions were analysed to broaden the understanding of the indexical meanings of her songs and language performance, to bring into focus not only the recipient but also the content and message of the music. Lily Allen's singing style was found to be characterised by an almost complete absence of American features, portraying, instead, British features with connotations ideal for expressing authenticity and a balanced rebelliousness within mainstream pop norms when offering the multi-faceted social critique characteristic of her songs.

Key words: style modification, Americanisation, Lily Allen, indexicality, Estuary English

1. Introduction

The quote in the title of this paper, derived from a YouTube comment [1] on Lily Allen's song 'The Fear' [2], seems to perfectly encapsulate Lily Allen as an artist. She achieved worldwide success with her debut album *Alright, Still* in 2006 and has since then released another 3, which have all earned commercial success and won numerous awards. One possible explanation for her immense success might be found in her unique singing style as she sings with a local variety rather than adopting the American sound, which Gibson notes is very 'common in popular music throughout (and beyond) the Anglophone world' [3, p. 3]. Building on Gerwin's research on Lily Allen's music, this paper sets out to provide a quantitative dataset based on an auditory analysis of her 6 most popular songs to corroborate the claim that Allen portrays British features in her singing and a lack of American variants. Through a multi-modal analysis of the content of her music, the paper, moreover, examines and discusses proposed motivations for her linguistic behaviour.

To lay the theoretical framework the paper will firstly present previous research on Americanisation in pop music and thereafter outline the features of the Estuary English variety.

2. Previous studies and theoretical framework

Trudgill's foundational study [4] marks the beginning of sociolinguistic research on Americanisation in British music. He set out to investigate British bands' notable digression from their spoken vernacular and the motivations for modifying their linguistic behaviour so. In describing this modified linguistic behaviour, he introduces the American pronunciation model consisting of 6

© 2024 Josephine Iben Kjeldbjerg. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

characteristic phonetic features that British singers adopt as they find them prototypically American.

In order to find a possible explanation to this phenomenon, he turns to Le Page's 'acts of identity' theory [4, pp. 252-253] as he argues that British musicians imitate the linguistic behaviour of the group that they want to identify with. Due to the African American origin of many genres within western pop music and Americans' consequent 'cultural domination' at the time, it was only appropriate, Trudgill argues, to converge one's sound to those who did it best [4, p. 254].

Le Page's theory also provides a notion of conflicting motivations, to which Trudgill assigns much salience. In a diachronic analysis of The Beatles' albums [4, pp. 260-261], he notices a decrease in American features and points to the weakening of the American model to explain this sudden shift. After the British Invasion and the global success gained by The Beatles, cultural domination had changed, and British music 'acquired a validity of its own' [4, p. 261].

Finally, he looks at the emergence of the punk movement in Britain. Here, he finds a motivation towards British covert prestige features and identification with the working-class youth conflicting with the continuing motivation towards the American model, reflected in hybrid accents [4, p. 262].

Building on Trudgill's research, Simpson offers a slightly altered framework to the principal model, which he labels 'the USA-5 model' [5, p. 345). Moreover, he brings a lot more focus to adoption of a specific persona and the topics of the lyrics so that the centre of attention is not only who is being sung to but also what is being sung about [5, p. 354].

Towards the end of his research, Simpson takes a look at the globalisation and centralisation of production in 90's pop music and the resulting perceived homogeneity, which was forcing bands to 'try to carve out their identity by searching for some generic label that marks out as different or unique' [5, pp. 361-362]. Among these labels he mentions indie and Britpop, which he sees as part of the explanation as to why the resonances and associations of the USA-model 'have altered inexorably over the years' [5, p. 364].

This is where Beal comes in with her influential research and qualitative language-ideological approach to the Indie-rock band, Arctic Monkeys [6]. Beal elaborates on Simpson's indications of change in associations of the USA-model as she brings in the theory of indexicality. When examining the mode of singing, it is highly relevant to look at indexicality, Jansen emphasises, since songs are nonspontaneous texts in which language becomes very agentive, intentionally manipulated 'to evoke certain associations and reactions' [7, p. 6]. Based on Eckert's notion of indexical fields, arguing that 'meanings of variables are not precise or fixed but rather constitute a field of potential meanings - an indexical field' [8, p. 453], Beal states that the associations of the USA-5 features have changed from being indexed as 'simply "American"' to 'mainstream pop' [6, p. 229].

The ubiquity of these features has resulted in an automatic employment, lacking any conscious act of identity, which is what certain British artists wish to dissociate themselves from by performing local varieties instead, indexing authenticity and independence from mainstream pop, Beal argues [6, pp. 224-25].

In a very recent study, Jansen [7] shifts the focus to audience perception as she argues that audiences and their evaluations play a decisive role in how social meanings become embedded and how singers modify their singing styles, especially now that the mediascape allows listeners to not just passively consume the music but also actively engage in discussions on language performances online [7, pp. 2-11].

In an interview-based research project, Jansen finds that especially the British participants are aware of American features, and generally criticise Americanisation in British music as it is seen to be artificial, mainstream and 'encroaching upon British music' [7, pp. 124-128]. The instances of British vernacular features evoke, by contrast, many positive attitudes among all participants as they are considered non-conformist and authentic by jeopardizing a potentially more widespread success for matters of local pride and cultural preservation [7, pp. 127-128].

3. A Closer Look on 'Estuary English' and Audience Responses

In her article on dialects in pop music [9], Gerwin claims that Allen is portraying Cockney features in her singing style. The indexical meanings attached to Cockney features at the time of Trudgill and Simpson, such as punk connotations of anti-establishment, are, however, enregistered differently in Allen's music as the features have become much more widespread and have given rise to a 'softer form', labelled Estuary English [9, p. 10].

The term, henceforth referred to as EE, was coined by Rosewarne in 1984 and is described as a variety widely distributed in the South-East of England, linguistically seen as a compromise and as a neutral 'middle ground' on a continuum with the overt prestige RP at one end and the local Cockney at the other [10, p. 3]. Rosewarne argues that the acquisition of EE is a two-way process of accommodation with different varieties converging as it lets speakers of local varieties accommodate upwards to sound more 'sophisticated' and educated while RP speakers accommodate downwards to increase 'street credibility' and to avoid 'posh' connotations [10, p. 7].

In this way, EE comprises features of both Cockney and RP and thus borrows 'the positive prestige from both accents without committing itself to either', Altendorf points out [11, p. 3]. This results in a 'vagueness' that makes the variant very difficult to linguistically demarcate. In spite of these complications in defining the variant, 'the term and concept of EE have caught on', she argues. In Well's consequent research on finding some phonological distinctions [12], Altendorf highlights three 'key' linguistic features, /t/-glottalisation, /l/-vocalisation and *th*-fronting. In her empirical research based on these features, she concludes that word-final /t/-vocalisation and /l/-vocalisation serve to mark EE off from RP whereas intervocalic /t/-glottalisation and *th*-fronting serve to marker off EE from Cockney. Among other consonantal features that mark EE off from Cockney, Wells also suggests H-dropping [12].

A majority of positive responses to her usage of EE become evident in the comment sections of her YouTube videos. A comment on the music video of her song 'Smile' notes that, '[m]ost ppl with English accents usually sound American, I'm glad her accent stays intact when she sings' [13], and is just one of many similar comments that support Jansen's argument that the listeners are aware of the general Americanisation and highly welcome the British features. Another comment on the music video of 'The Fear', stating that 'it's so refreshing to hear someone sing in their own accent! The twangy American accent lends itself well to lyrical expression, but accents are many and varied; let's start making more music with individuality and sincerity' [14], also shows awareness of possible motivations for Americanisation and confirms the suggested associations of authenticity in local varieties.

4. Methodology

As Gerwin did not provide any analysis of Allen's singing and only touched briefly on one song, I wanted to build on her claim and provide a quantitative dataset as well as a broader multi-modal analysis of the content of her songs to ascertain that she does portray both Cockney and RP features and an absence of American variants and to help clarify the motivations behind this portrayal. Thus, a quantitative auditory analysis is conducted of her top 6 songs on Spotify [15, 16, 17, 18, 19, 20], listed in Table 1 below.

The analysis is based on the USA-5 features and the EE features mentioned above. These include the 5 American variables: the TRAP, LOT, and PRICE vowels; intervocalic /t/, and postvocalic /r/ and the 4 Cockney variables: word-initial /h/, word-or syllable initial TH, intervocalic and final /t/, and preconsonantal and prepausal /l/. The American vowels are only examined in stressed syllables to avoid unstressed variants. The analysis is done by identifying the possible occurrences of the linguistic variables in the lyrics and then analysing how these are realised. All lyrics except from backing vocals are included in the analysis. The results from each song are added together to obtain a percentage of how often the American and Cockney variants occur. In the succeeding section, non-linguistic dimensions in the 6 songs will be analysed in a multimodal approach, analysing the lyrical content and the music videos.

5. Results

The results of the auditory analysis of Allen's use of the USA-5 variables are illustrated in Figure 1 and Table 1 and the 4 Cockney variables are illustrated in Figures 2 and Table 2. Figure 1, portraying the USA-5 model, shows that there is an almost complete absence of the American variants in Allen's top 6 songs with one exception of the PRICE vowel. Table 1 shows that the American monophthongisation of the PRICE vowel, however, is centralised in only one song, namely 'Smile'. In this song, the monophthongised PRICE vowel only occurs in the two words *smile* and *while* which do, however, appear frequently in the chorus and are consequently repeated four times, making up the 22 out of 141 possible instances in the 6 songs.

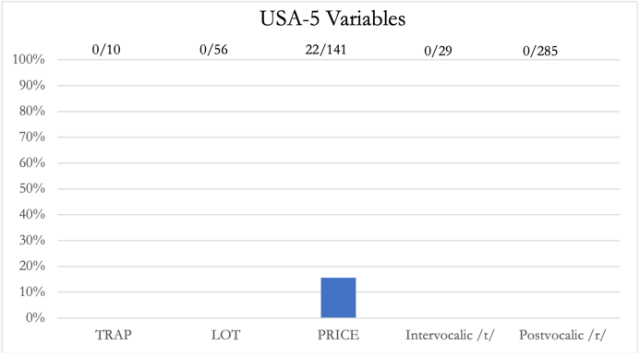


Figure 1: *percentage of USA-5 variables in the 6 songs.*

Table 1: *percentage of USA-5 variables in the 6 songs.*

USA-5	TRAP	LOT	PRICE	Intervocalic /t/	Postvocalic /r/
LDN (2006)	0/1	0/3	0/63	0/10	0/33
Smile (2006)	0/1	0/4	22/31	0/4	0/25
Not Fair (2009)	0/0	0/19	0/11	0/4	0/60
Fuck You (2009)	0/2	0/7	0/14	0/4	0/39
The Fear (2009)	0/1	0/16	0/11	0/2	0/38
Hard Out Here (2013)	0/5	0/7	0/11	0/5	0/90
Total	0/10	0/56	22/141	0/29	0/285
Percentage	0%	0%	15.6%	0%	0%

Figure 2, which portrays the Cockney variables, displays a slightly more complicated pattern. As proposed by Wells, the first variable, word-initial /h/, is not dropped in any of the 101 possible instances but is realised as in RP. Another variable also marking EE off from Cockney, the word- or syllable- initial TH, is realised like in RP in 42 of the 44 possible instances. The two remaining instances are, however, realised like in Cockney as the labiodental fricative, [v], which, as seen in Table 2, only occurs in her song 'Not Fair'. This pronunciation appears in the words *with* and *other* and portrays inconsistency in Allen's singing

© 2024 Josephine Iben Kjeldbjerg. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

as these words also appear in her songs 'Smile', 'The Fear', and 'LDN' in which they are realised as the voiced dental fricative, [ð].

The third variable, intervocalic and word-final /t/, shows that Allen portrays the Cockney variant in 177 out of 204 possible instances. Despite being a feature proposed to separate EE from Cockney, Allen glottalises all instances of intervocalic /t/ except from the one inconsistent word, *matter*, in 'The Fear'. The rest of the instances where Allen does not glottalise are word-final /t/. These instances are only seen, however, in environments where the variable is followed by either a /t/, /ð/ or /d/, resulting in assimilation of the word-final /t/. This is evident in the five instances of *want to* found in 'Not Fair', 'Fuck You' and 'Hard Out Here' where the word-final /t/ is realised as [t]; in the four instances of *at the* found in 'Smile' and 'The Fear' where the word-final /t/ is realised as [ð]; and in the one instance of *it doesn't* found in 'LDN' where the word-final /t/ is realised as [d]. Allen is otherwise consistent in glottalising the word-final /t/ in all other instances.

The last variable, preconsonantal and prepausal /l/, is slightly less consistent as only 95 out of 131 possible instances are being vocalised in spite of being a suggested feature of EE. However, most of these instances, as seen in Table 2, are centralised in her song 'Smile' where the two words *smile* and *while* make up 2/3 of the instances in all six songs. Among the remaining instances are for example the word *middle* from 'Not Fair' and the three instances of *well* in 'Fuck You' and 'Hard Out Here' but these only represent infrequent deviations of the otherwise high level of /l/-vocalisation found in the other five songs.

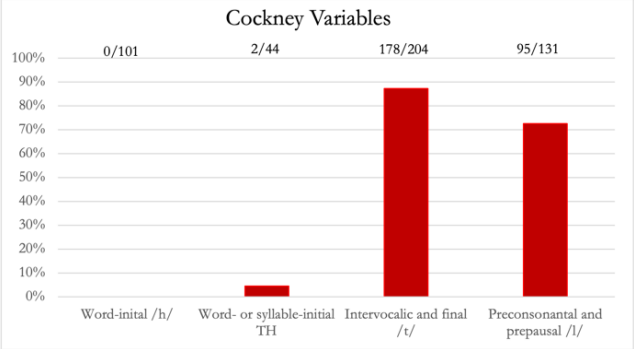


Figure 2: *percentage of Cockney variables in the 6 songs.*

Table 2: *percentage of Cockney variables in the 6 songs.*

Cockney	Word-initial /h/	Word- or syllable-initial TH	Intervocalic and final /t/	Preconsonantal and prepausal /l/
LDN (2006)	0/10	0/13	28/29	21/22
Smile (2006)	0/8	0/3	36/43	21/45
Not Fair (2009)	0/9	2/15	25/31	14/15
Fuck You (2009)	0/15	0/3	32/33	15/18
The Fear (2009)	0/6	0/8	22/30	16/21
Hard Out Here (2013)	0/53	0/2	35/38	8/10
Total	0/101	2/44	178/204	95/131
Percentage	0%	4.54%	87.25%	72.51%

6. Discussion

A larger sample of Allen's songs would naturally be necessary to give a comprehensive picture of the degree of mixing Cockney and RP features in her singing style, but the above analysis does indicate that features from both British varieties are present and noticeably shows that there is an almost complete absence of the USA-5 features in her top 6 songs.

To understand the notable deviation from the general pattern in her song 'Smile', it is beneficial to consider Simpson's notion of the influence of the 'mode of discourse' [5, p. 360]. He points out that 'the less a singer "sings"', the more the singing style will portray the singer's own vernacular sound as the mode of speaking is not influenced by the same physiological demands. Just as Arctic Monkey's Alex Turner's singing style is more characterised by 'melodic storytelling' compared to British singer Adele's more traditionally melodic style [7, p. 26], Allen also portrays elements of a slightly faster, more speech-like delivery that allows for the British variants to come through. In the chorus of 'Smile', however, the words *while* and *smile* are temporally prolonged, and the tones are changing in a descending pitch movement, rendering the deliverance of the words more melodic. This explains the unusual occurrence of the monophthongised PRICE vowel since this American feature, as Coupland notes [21, p. 296], is 'easier to hold' for longer than the equivalent British diphthong.

Whether the personal inconsistencies seen in the *th*-fronting in 'Not Fair' are a case of Allen style-shifting when singing, performing more Cockney features, is difficult to say but would be interesting to examine in a comparison with her speech. The reasons behind the inconsistencies between Altendorf and Wells' suggested features of EE and Allen's lower degree of /l/-vocalisation and higher degree of intervocalic glottalisation are difficult to clarify due to the complicated linguistic demarcation of EE. Owing to these difficulties in defining a uniform variety, this paper's intention is merely to ascertain that Allen portrays a mix of Cockney and RP features and a lack of American variants. So why would Allen choose to portray the Estuary English variant in her music and almost completely leave out any American features? Unlike Arctic Monkeys, who represent the desire of the indie genre to be seen as independent from the 'global marketing machine' [6, p. 230], Allen is a mainstream pop artist who has worked with major record labels to promote her music. Her divergence from the Americanisation is thus not a case of genre appropriateness. Nor do earlier suggested influencing factors for style-shifting such as musical role models, audience accommodation and commercial success [22, p. 117] seem to be relevant for Allen's singing style as she has been performing and touring in the USA since her first album without changing her sound. Seeing that Allen is not influenced by genre and who she is singing to and for, it

seems appropriate to rather focus on the 'constraints of topic and field' as Simpson proposes [5, p. 354].

Gerwin points out that Allen herself has expressed that she has intended 'to do some social commentary and comment on pop culture' [9, p. 11] through her music. As a woman coming from a middle-class London background, the Estuary features become the perfect compromise in offering social commentary authentically and for 'expressing female identity a bit rebelliously, with a charming, not too outrageous, or provocative-seeming directness', Gerwin argues [9, p. 11]. This is encapsulated in Allen's song 'LDN' in which she highlights the high level of social inequality and crime in London, 'distancing herself from the glorification and hype surrounding the capital', Gerwin points out. In this way, the Cockney working class features in her singing increase Allen's authenticity in delivering the criticism compared to a usage of the unrelated American sound or a pure 'posh' RP sound with potential condescending connotations. She simultaneously benefits from a certain degree of 'educatedness' when also portraying some RP features.

To get a more comprehensive understanding of the indexical meaning of Allen's music and whether this social commentary is generally present in her songs, the following section further examines the content of her top 6 songs in a multi-modal analysis of her lyrics and music videos.

7. Multimodal Content Analysis

7.1. British cultural references

To accentuate her performed British persona, Allen has included a number of British lexical variants in her lyrics. Besides referencing London in the title of 'LDN', she also names 12 smaller areas within London. Allen also uses the word *slapper* in the song, which is defined in *Bloomsbury Dictionary of Contemporary Slang* as a 'working class term from East London and Essex' [23, p. 468]. In her song 'The Fear', she makes a reference to two of UK's most popular tabloid newspapers in the line 'I'll look at the The Sun and I'll look in The Mirror' [24], which is a word play that most likely will get recognised by most British people but might go unnoticed by non-Brits. Another British cultural reference is seen in her song 'Smile' in which she sings 'with a little help from my friends' [25], which very likely is an intertextual reference to The Beatles' song by the same name, drawing on the Brits' pride in them.

7.2. Social commentary in Allen's music

Like 'LDN', the rest of Allen's top 6 songs also seem to offer a direct social critique. Beginning with her song 'The Fear', Allen offers a sarcastic critique of false images presented on social media and modern society's consumerism: 'And I am a weapon of massive consumption/And it's not my fault, it's how I'm programmed to function' [24]. Political critique is also seen clearly in her song 'Fuck You', which she has publicly proclaimed is addressed to former president George W. Bush [26], as she presents anti-homophobia and anti-war sentiments, singing 'Your point of view is medieval' [27]. The political message of the song was also reaffirmed in 2022 as Allen performed the song with Olivia Rodrigo as a response to the Supreme Court's decision to eliminate the constitutional right to abortion in the US [28].

In the rest of the songs 'Smile', 'Not Fair' and 'Hard Out Here' there seems to be a growing feminist agenda. Beginning at the somewhat innocent level, 'Smile' portrays an empowering story of a woman turning her pain and anger towards a cheating ex-boyfriend into a satisfaction of seeing him desperate and alone as she sings 'At first when I see you cry/Yeah it makes me smile' [25]. Increasing the intensity, 'Not Fair' criticises the more general societal problem of inequality in men and women's sex lives as it portrays men as often incapable, or perhaps even unwilling, to sexually satisfy women seen in the line 'Oh, you're supposed to care/But all you do is take' [29].

Finally, in 'Hard Out Here', a full-scale 'feminist anthem' according to a review in *Rolling Stone* [30], becomes evident. Allen sheds light on old gender roles, objectification, and twisted body ideals of women, and she reappropriates the sexist slur *bitch* [31], which can be used as a tool 'to challenge the patriarchal and misogynist views' [32]. In the characteristic line 'Forget your balls and grow a pair of tits' [31] she inverts the phrase 'grow a pair' to be associated with breasts, thus ascribing strength to being a woman and prompting men to put themselves into the place of women.

7.3. Music videos

Looking at the content of the lyrics is one factor that needs to be accounted for, but Bell and Gibson [33, p. 566] also point to the importance of the visual modality of music videos. This importance is seen in the music video of 'LDN' [34] whose general theme is showing a contrast between a glorified version of London and the reality of urban problems behind the false façade. This is exemplified in the segment where she sings about people having an 'al fresco' lunch in the park in the same moment as the colourful pictures fade into dull colours portraying a homeless pair eating outside with no other place to go, which adds more sarcastic depth than the lyrics on their own.

Examining the music video of 'Hard Out Here' [35] also supports and adds new meanings. In the video's intro, Allen is lying on a plastic surgeon's table being criticised for her body's appearance after having given birth to two children, which certainly supports the lyrics' critique of the female body ideal. The video also adds an intertextual reference, otherwise missed, to Robin Thicke's music video for his song 'Blurred Lines' [36], which was highly criticised in contemporary articles for being a 'rape anthem' [37] and 'completely misogynist, objectifying naked women' [38].

© 2024 Josephine Iben Kjeldbjerg. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

This is seen as Allen incorporates a scene with silver balloons spelling 'Lily Allen has a baggy pussy' as a direct response to Thicke's identical scene where his balloons spell out 'Thicke has a big D'. In this way, 'Hard Out Here' also becomes a critique, more specifically, of misogyny in the music industry, also portrayed in the seemingly hypocritical element of half-naked women dancing in a sexualised manner, matching the sarcasm of the other aspects of the song.

8. Conclusion

This study has analysed a small sample of Lily Allen's most popular songs in order to provide a quantitative dataset for the claim that Allen portrays Estuary English in her singing style. The paper has argued that Allen's singing style is characterised by an almost complete absence of Americanisation and contrarily includes Estuary features. Since it encompasses both features of RP and Cockney, Estuary English becomes ideal for Allen to express a balanced rebelliousness when offering the multi-faceted social critique characteristic of her songs. Instead of commenting from the outside by for instance producing indie-rock music like the Arctic Monkeys, Allen chooses to criticise the mainstream pop culture from within, exploiting the commercial format to disseminate her ideas and values. The applied multimodal approach has shown how interdisciplinary supplements can corroborate and broaden the sociolinguistic understanding of Lily Allen's music.

9. References

1. Impas, J. (2020). *She's Like That Rebellious Girl Who Doesn't Try Too Hard.* [Comment on the video 'Lily Allen | The Fear (Official Video – Explicit Version)']. *YouTube.* https://www.youtube.com/watch?v=q-wGMlSuX_c
2. Allen, L. (2009). The Fear. On *It's Not Me, It's You* [Audio file]. Retrieved from https://open.spotify.com/track/7lIyBKp2KX7ABDjDfo9EtL?si=b9ea5225a84a4404
3. Gibson, A. (2023). 'Pop Song English as a Supralocal Norm'. *Language in society*, pp. 1–28.
4. Trudgill, P. (1997). Acts of Conflicting Identity: The Sociolinguistics of British Pop-song Pronunciation. In Coupland, N & A. Jaworski. *Sociolinguistics, A Reader and Coursebook* (pp. 251-65). New York: Palgrave.
5. Simpson, P. (1999). Language, Culture and Identity: with (Another) Look at Accents in Pop and Rock Singing. *Multilingua, 18* (4), pp. 343-367.
6. Beal, J. C. (2009). 'You're Not from New York City, You're from Rotherham': Dialect and Identity in British Indie Music. *Journal of English Linguistics, 37* (3), pp. 223-240.
7. Jansen, L. (2022). *English Rock and Pop Performances: A Sociolinguistic Investigation of British and American Language Perceptions and Attitudes.* Amsterdam/Philadelphia: John Benjamins Publishing Company.
8. Eckert, P. (2008). Variation and The Indexical Field. *Journal of Sociolinguistics, 12* (4), pp. 453-476.
9. Gerwin, J. (2017). The Beatles, Arctic Monkeys and Lily Allen – Dialects and Identity in British (Indie) Pop Music. In Anderwald, L. & J. Hoekstra. *Enregisterment – Zur Symbolischen Bedeutung Sprachlicher Variation*, (pp. 47-65). Frankfurt: Peter Lang.
10. Rosewarne, D. (1994). Estuary English: Tomorrow's RP? *English Today, 10* (1), pp. 3-8.
11. Altendorf, Ulrike. (1999). 'EJtuary English: Is English Going Cockney?'. *Moderna Språk 93* (1), pp. 2-11.
12. Wells, J.C. (1994). 'Transcribing Estuary English: a discussion document'. *Speech Hearing and Language: UCL Work in Progress 8*, pp. 259-267.
13. 10199ULTRAMATIC. (2022). *Most Ppl with English Accents Usually Sound American.* [Comment on the video 'Lily Allen | Smile (Official Video)']. *YouTube.* https://www.youtube.com/watch?v=0WxDrVUrSvI
14. TheImmutableTruth. (2016). *It's so Refreshing to Hear Someone Sing in Their Own Accent.* [Comment on the video 'Lily Allen | The Fear (Official Video - Explicit Version)']. *YouTube.* https://www.youtube.com/watch?v=q-wGMlSuX_c
15. Allen, L. (2006). LDN. On *Alright, Still* [Audio file] Retrieved from https://open.spotify.com/track/5wD6de7Fz1QfxVNkv1SEy7?si=e7ab6fca01b6418a
16. Allen, L. (2006). Smile. On *Alright, Still,* [Audio file] Retrieved from https://open.spotify.com/track/79z9QkhHePTFsSeVw9uyj0?si=9bf4c3d8002b4a98
17. Allen, L. (2009). Not Fair. On *It's Not Me, It's You* [Audio file] Retrieved from https://open.spotify.com/track/5st5644IlBmKiiRE73UsoZ?si=22e55e0b07b547a4
18. Allen, L. (2009). Fuck You. On *It's Not Me, It's You* [Audio file] Retrieved from https://open.spotify.com/track/5st5644IlBmKiiRE73UsoZ?si=fe34221a44c64b46
19. Allen, L. (2009). The Fear. On *It's Not Me, It's You* [Audio file] Retrieved from https://open.spotify.com/track/7lIyBKp2KX7ABDjDfo9EtL?si=1ff6e16064724ffd
20. Allen, L. (2014). Hard Out Here. On *Sheezus* [Audio file] Retrieved from https://open.spotify.com/track/6FzUX4kFBss56SSjAsLi7t?si=e172a203488f47b1
21. Coupland, N. (2009). 'The Mediated Performance of Vernaculars'. *Journal of English Linguistics 37* (7), pp. 284–300.
22. Jansen, L. (2018) 'Britpop Is a Thing, Damn It: On British Attitudes Toward American English and an Americanized Singing Style'. In *The Language of Pop Culture*, edited by Valentin Werner, 116-135. 1st ed. Routledge.
23. Thorne, T. (2014). *Dictionary of Contemporary Slang.* 4. ed. London: Bloomsbury.

24. Allen, L. (2009). The Fear [Lyrics]. Retrieved from https://genius.com/Lily-allen-the-fear-lyrics
25. Allen, L. (2006). Smile [Lyrics]. Retrieved from https://genius.com/Lily-allen-smile-lyrics
26. *New Musical Express*. (2009, April 3). Lily Allen gets too political in LA. https://web.archive.org/web/20090406105331/https://www.nme.com/news/lily-allen/43881
27. Allen, L. (2009). Fuck You [Lyrics]. Retrieved from https://genius.com/Lily-allen-fuck-you-lyrics
28. Nugent, A. (2022, June 25). Olivia Rodrigo brings out Lily Allen to sing 'F*** You' in response to Supreme Court over Roe vs Wade. *The Independent*. https://www.independent.co.uk/arts-entertainment/music/news/olivia-rodrigo-glastonbury-lily-allen-b2109446.html
29. Allen, L. (2009). Not Fair [Lyrics]. Retrieved from https://genius.com/Lily-allen-not-fair-lyrics
30. Blistein, J. (2013, November 12). Lily Allen Returns With a Smirk in 'Hard Out Here'. *Rolling Stone*. https://www.rollingstone.com/music/music-news/lily-allen-returns-with-a-smirk-in-hard-out-here-241643/
31. Allen, L. (2014). Hard Out Here [Lyrics]. Retrieved from https://genius.com/Lily-allen-hard-out-here-lyrics
32. Koponen, I. (2021). 'I love being a bad bitch': Reappropriation and Indexicalities of Bitch by American Female Singers on Twitter. *HELDA,* pp. 1-68.
33. Bell, A., & Gibson, A. (2011). Staging Language: An Introduction to the Sociolinguistics of Performance. *Journal of Sociolinguistics, 15* (5), pp. 555-572.
34. Allen, L. (2008, November 27). *LDN* [Video]. YouTube. https://www.youtube.com/watch?v=wmYT79tPvLg
35. Allen, L. (2013, November 12). *Hard Out Here* [Video]. YouTube. https://www.youtube.com/watch?v=E0CazRHB0so
36. Thicke, R. (2013, March 20). *Blurred Lines* [Video]. YouTube. https://www.youtube.com/watch?v=yyDUC1LUXSU
37. Koehler, S. (2013, September 19). From the Mouth of Rapists: Lyrics of Robin Thicke's 'Blurred Lines'. *Pacific Standard*. https://psmag.com/social-justice/mouths-rapists-lyrics-robin-thickes-blurred-lines-66569
38. Bein, K. (2013, August 22). Robin Thicke's 'Blurred Lines' Is About Rape and Seven Other Songs You Misunderstood. *Miami New Times*. https://www.miaminewtimes.com/music/robin-thickes-blurred-lines-is-about-rape-and-seven-other-songs-you-misunderstood-6436012

© 2024 Josephine Iben Kjeldbjerg. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-007

The good, the bad, and the funny Russian: linguistic variation in telecinema

Freja Hovgaard Knudsen[1]

[1]Department of English, Aarhus University, Denmark

201804725@post.au.dk, 03maj97@gmail.com

ABSTRACT

Movies are filled with linguistic variation. However, the speech in movies is not spontaneous and authentic and the linguistic varieties found therein do not necessarily rely on the vernacular of a real person. Instead, they rely on indexes to communicate information about the characters to the audience. On the surface, this information seems to merely pertain to the factual background of the character, e.g., their native language. This paper argues that this is not the case. Focusing on Russian English, this paper investigates the usage and implicit indexes of a specific feature of Russian English, namely article omissions. Quantitative data for articles and article omissions from eleven Russian characters from four movies have been collected and analysed using R. Statistical testing suggests the number of article omissions depends on whether the character is a comic relief character, with a positive correlation between being a comic relief character and article omissions. The number of articles depends on whether the character is a protagonist and the year of release, with a positive correlation between articles and the year of release and articles and being a protagonist. Qualitative observations support these findings.

Key words: indexicality, telecinema, Russian English, sociolinguistics, stereotyping, action telecinema, linguistic variation

1. Introduction

Linguistic variation is one of the tools telecinematic productions can utilize to provide characterization for the fictional people living within the telecinematic universe [1]. Within the intradiegetic boundaries of a telecinematic universe, the characters are 'real'. As such, the information communicated to the audience through the characters' speech productions is likely at least partially motivated by a desire to mimic a 'realistic' correlation between the linguistic productions of a character and the diegetic background and history of said character [1; 2; 3]. The result is that e.g. a character from Russia should speak English influenced by Russian features since such cross-linguistic influence is commonly observed in real life [4]. This is true even if the actor giving voice to the character does not have Russian as their L1.

However, linguistic variation in telecinema is not solely guided by real-life occurrences of language use [1; 2], instead imitations of real-life dialects often also draw on linguistic stereotypes [5], thereby communicating non-factual embedded information about the characters to the audience.

This paper aims to examine the English dialogue of eleven Russian characters from a selection of telecinematic action productions to investigate potential indexes of features of Russian English. Russian characters, especially those found in action telecinema, are productive for such investigations as they historically have been used to fulfil the roles of antagonists [6; 7; 8] and only more recently are being given other plot functions and have been provided with more varied characterizations. This allows the categorization of Russian English characters into more than one group which is necessary when wishing to investigate the indexicality of certain features without a perception study. Differences in the distribution of features between groups signal the indexes of said features.

English produced by native Russian speakers is marked by several different features [9], all contributing to a Russian English dialect. Either individually or in combination, these features index Russianness and therefore all have the potential to be used in telecinematic portrayals of Russian characters. However, these individual features may index more than simply nationality. Therefore, careful attention should be paid to whether there is homogeneity in the distribution of these features among different types of Russian English characters since a lack of homogeneity could indicate that certain linguistic features index specific character traits or plot functions in telecinema. The paper will investigate the article omissions of eleven Russian characters from four telecinematic productions.

© 2024 Freja Hovgaard Knudsen. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

1.1. Hypotheses

Casual observation suggests that comic relief is becoming increasingly more typical in action telecinema. As article omissions are also seemingly becoming more typical for Russian characters in action telecinema, it could be an indication that the two are correlated and perhaps even causally linked, i.e. that article omissions index comic relief.

The paper hypothesises that article omissions are used for comic relief and are primarily seen with a certain subtype of character whose main function in the telecinematic narrative is to provide comic relief, with comic relief defined as 'a storytelling plot device that's meant to lighten the project's mood and keep the viewer entertained' through 'pieces of dialogue, actions and/or characters' [10]. The most striking examples of comic relief characters are those whose function in the story, apart from the provided comic relief, could be filled by inanimate objects or natural phenomena.

2. Literature

Studying linguistic variation in telecinema is important as language in telecinema is ideologically potent [11]. It utilizes established language conventions *and* affects the audience's attitude towards varieties [6]. The stylized speech in telecinema relies on and affects indexical information. Furthermore, many viewers are only exposed to certain dialects and types of characters through telecinema which means that the underlying implications of linguistic variation in telecinema are likely to influence real-life attitudes [6].

Considering Russian characters in telecinema, it is important to note that 1.) Russian characters are often played by non-Russian actors and 2.) many people are unlikely to encounter Russian English anywhere else than in telecinema. This means that 1.) since Russian is not the actor's L1, inclusions of various Russian English features in the character's speech productions are by design rather than a natural occurrence of cross-linguistic influence and 2.) many audience members will not have any other reference as to what Russian English entails or the behaviour and cultural norm of a Russian person. Therefore, the portrayal of Russians and Russian English in telecinema must in the production stages rely on existing indexes and upon release create or reinforce indexes. This means that a key term for understanding the importance of linguistic variation, e.g., Russian English, in telecinema is indexicality.

2.1. Indexicality

Indexicality can be described as 'the complex association between linguistic features and social meaning' [1, pp. 18-19], and the performance medium especially 'provides a forum where an audience's attention can be drawn to indexical relationships, reinforcing some social meanings and reinterpreting others' [12, p. 561]. In telecinema, these indexical relationships are created through the combination of the storyline, the visuals (cinematography, mise-en-scène, etc. [1; 3; 13]), music, and speech productions. While focusing on the indexes of linguistic features, it is important to keep in mind that the other aspects of a character interact with how the audience is prompted to regard the character. As such telecinema relies on genre conventions and already established indexes, but it also aids the process of enregisterment, i.e. the 'process by which a style becomes engraved in the public mind as indexing certain social personas or "characterological figures"' [12, p. 561].

Traditionally, Russian characters in Western-produced action telecinema have fulfilled the role of villain or "bad guy" [2; 8]. Russian English, or at least certain realisations of Russian English, have likely indexed 'evilness'. Seemingly in recent years, Russian characters have begun to fulfil other roles than that of the villain - e.g., as comic relief characters, helping the protagonists. As such the indexicality of features typical of Russian English could be changing or expanding. This is possible since indexes are not fixed; they shift and change over time, and any use of a *n*-th order indexicality is always potentially creating a new *n+1*st order indexicality, a 'competing structure of values potentially indexed in-and-by a communicative form of the *n*-th order, depending on the degree of intensity of ideologization' [14, p. 194]. As such indexes are not unambiguous, meaning multiple indexes can co-exist at the same point in time with varying familiarity for listeners [15]. It is possible that some features of Russian English still index 'villainous intent' while others index humour.

Generally, some indexes might be more prominent than others with mass media and pop culture acting as main distributors of index familiarity and prominence as they quickly reach a much greater and more diverse audience than most other linguistic productions do [1; 16]. This is especially true of the most popular telecinematic genres, one of which is the action genre [17]. The linguistic variation of such telecinematic productions should therefore be carefully examined.

2.2. Stylisation

Speech in telecinema can be categorised as stylised speech. Stylisation relies on the indexicality of various linguistic varieties to convey meaning as it is 'the mannered adoption of another's voice' [12, p. 560] that 'brings into play stereotyped semiotic and ideological values associated with other groups, situations or times' [16, p. 350]. In telecinema, the goal is for the actor to 'become' the character in the eyes (and ears) of the audience whereas stylisation in everyday speech is of shorter duration and has a greater distance between speaker and voice, communicating more explicit "'as if I endorsed what this voice says'"-statements [16, p. 349]. Even so, stylisation 'may not be fundamentally different in mediated and nonmediated contexts' [16, p. 351].

One important distinction between stylised speech in everyday contexts and speech in films is that speech in telecinema is 'scripted, written and rewritten, censored, polished, rehearsed, and performed' and even improvised lines have been 'spoken by impersonators, judged, approved, and allowed to remain' [18, p. 18] with even the vernacular intentionally reproduced, not naturally occurring [12]. Arguably, this only strengthens the case that the indexes of movie speech should be paid attention to

since the speech that reaches the audience has been carefully crafted and selected at multiple stages.

3. Data and Methods

3.1. Data

The English dialogue of eleven Russian characters from four telecinematic productions, released between 2015 and 2022, has been examined with a focus on article omissions. Three out of the four productions are installations in widely recognised pop-cultural franchises, namely the *Marvel Cinematic Universe*, the *Stranger Things* Netflix sensation, and the *Bond* franchise. As pop culture phenomena, these telecinematic productions and the linguistic realisations found therein have 'the potential to foster a particular worldview and therefore may fulfil an ideological function' [11, p. 3]. They therefore provide fertile ground for examining the indexes of linguistic features and dialects. The last of the four selected telecinematic productions does not have the same pop-cultural prominence as the other four. It does, however, have the advantage of containing an L1 Russian speaker which could prove beneficial in further research. As this production falls within the same telecinematic genre as the three pop cultural phenomena, it has been judged that the data provided by it is comparable to the other three. Furthermore, the eleven characters from the four telecinematic productions provide variation in character types and plot functions.

From *Black Widow* [19] the characters Natasha, Yelena, Alexei, Melina, and Dreykov have been chosen. Except for Dreykov, all four characters are on the side of the protagonist, with Natasha as the main character. The film depicts the end of a mission for Dreykov in the US, the subsequent separation of the four protagonists, and, many years later, their reunion and successful attempt at taking down Dreykov and his unethical training of young girls.

From *Stranger Things* Season 4 [20] the characters Antonov and Yuri have been selected. Antonov, a guard at a Russian prison, takes payment from American prisoner and hero, Hopper, to help orchestrate his escape. The agreement is that Yuri will provide transportation out of the USSR, but he betrays both Antonov and Hopper who are then imprisoned. Facing the main threat of the *Stranger Things* franchise, Antonov quickly becomes a traditional sidekick. Yuri is soon overpowered by Hopper's American friends and reluctantly dragged along to save him. At last, Yuri does what he was meant to do all along and flies the heroes back to the US. Plot-wise, the role of Yuri could be fulfilled by a natural phenomenon such as a storm, indicating he might be a comic relief character.

From *No Time to Die* [21], the main villain Safin and his helper Obruchev have been chosen. Safin is a classic villain without any notable redeeming values except perhaps his tragic backstory. Throughout the movie, he is seen both personally killing others and orchestrating weapons of mass destruction. Obruchev is a scientist working undercover for Safin. He plays a key role in the development of the weapon and supports Safin's vision, but he is also often the butt of the joke and is not equipped to take part in any kind of physical fight. In many scenes, Obruchev's presence with the suitcase containing the weapon could be replaced with the suitcase on its own which indicates that his role could be that of a comic relief character.

From *The Man from U.N.C.L.E.* [22] one of the main characters, Illya, as well as his commander from the KGB, Oleg, have been selected. Oleg primarily communicates with Illya in Russia and only speaks Russian in one scene. Despite the tension between Russia and America, Oleg orders Illya to work with American CIA agent Solo – but to maintain a willingness to ultimately betray and kill Solo. Gradually, however, Illya and Solo become close. At the end of their mission, Illya defies Oleg's direct orders when he does not kill Solo.

This selection of characters ensures a variety of Russian characters, including protagonists, antagonists, and characters providing comic relief.

3.2. Methods

The number of articles and article omissions for each character has been found through transcriptions of the scenes involving relevant characters. Only articles that are normally explicit have been counted (both when present and when omitted). This means that e.g. the implicit indefinite plural articles have not been counted. Neither have other types of determiners such as e.g. *my*, *his*, and *that*.

To examine the number of article omissions in relation to the hypothesis that article omissions index comic relief, each character has been defined as either being a comic relief character or not. While the designation of a character as either a comic relief character or not is subjective, a few guidelines have been utilized in the categorisation: Firstly, do their scenes make people laugh; secondly, are they the reason for said laughter or is another character the primary force behind the humour; thirdly, are they an important emotional focal point? If so, they are likely not a comic relief character; and lastly, the "cartoon-character" consideration: If something went wrong in one of their scenes, e.g. stepping on a creaky floorboard while needing to be completely quiet, would they utter an 'oh-oh' in the following close-up, similar to the reaction of cartoon characters falling from great heights, or would they remain silent? This categorisation situates Alexei, Yuri, Obruchev, and Illya as comic relief characters. Considerations have also been made as to whether Yelena should be categorised as such, but ultimately it was decided that she is too much of an emotional focal point in *Black Widow* to be categorised as a comic relief character.

To control for other variables, notes have been taken for each character as to whether the character is a protagonist or an antagonist, their gender/sex, the year of the movie release, and the L1 of the actor portraying the character. The L1 of the actor is registered based on the nationality of the actor as indicated on Wikipedia and IMDb.

© 2024 Freja Hovgaard Knudsen. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

3.3. Russian English

Zavyalova et. al. [9] provides a useful guideline for which features are noteworthy when examining inauthentic, stylised, movie-Russian English:

Phonetically, English contains both vowel and consonant phonemes that are not present in Russian, making it common for Russians to replace English sounds with a perceived similar Russian sound. The result is a lack of discrimination between phonological contrasts such as [θ]/[s], [θ]/[f], [ð]/[z], [ð]/[d], [dʒ]/[tʃ], [ŋ]/[n], [w]/[v], [ɪ]/[i], and [æ]/[e].

Furthermore, Russians speaking English tend to devoice word-final voiced consonants, pronounce silent letters, roll their /r/s, and produce more velarised /h/-sounds, making them sound 'harder' [9, p. 85] than what is usual for native English speakers.

On the prosodic level, Russian as a first language wields a significant effect on English productions, e.g., through resyllabification in places where the vowel sound otherwise has been deleted. Especially the use of Russian stress-patterns, rhythm, and intonation 'gives the Russian speaker of English away' [9, p. 92].

Grammatically, Russian English tends to be 'characterized by hypercorrect … grammar, with full forms, lack of elision and assimilation' [9, p. 99], use of non-standard tense, and omissions of articles.

Syntactically, Russian English speakers tend to use a looser word order than what is typical for English. Russian English is also marked by complement loss after certain verbs and special types of tag questions [9].

Generally speaking, performed speech is reductive, meaning that it 'has a reduced amount of, or is even free from, "disruptive" elements such as false starts, repetitions, self-corrections, etc' [11, p. 562; 23], so attention should be paid to any included 'disruptive' elements, and for which characters these elements have been allowed. 'Non-standard' features such as those common for Russian English can be regarded as 'disruptive' elements and it should therefore be considered which such features are included and when and for whom they are utilised, since actual linguistic outcomes in telecinema are 'strongly determined by production (e.g., planning and editing of performed texts) and reception (e.g., asynchronicity) circumstances' [11, p. 568], i.e. the so-called 'performance filter'.

When a language variety, such as e.g. Russian English, is used in performances, it is typical that some features of the variety will be utilised, while others are omitted [2]. Among the features used, 'the characteristics or frequency of features of the targeted variety' will be both under- and overshot [12, p. 568]. Furthermore, mis-realisations of features are also likely to occur 'perhaps intentionally, or drawing on stereotypes, or through incapability' [12, p. 568].

Lawless (2014) found Russian action characters within the *Bond* franchise in the period 1962-2012 to realise /r/ 'fully and strongly sounded' as well as do a 'strong pronunciation of sound [ŋ] at the end of the word; soft [l], no aspiration of sounds [d] and [t]; pronunciation of sound [z] or [s] instead of [ð] and [θ] and the use of sound [v] instead of [w]' [6, p. 88]. For animated films, studies of Russian characters have described the specific features used to convey Russianness in less detail instead concluding that the 'function of this often exaggerated and contrived Russian-accented English is to provide humour and to contrast with other characters' [7, p. 199].

4. Results

The data of each character is shown in Table 1. Their total amount of articles and omissions are shown in Table 2.

It should further be noted that Antonov has two omitted determiners: 'I can tell by voice that she's very pretty' [20, episode 3, season 4] and "I change mind" [20, episode 4, season 4)). Alexei has one omitted determiner: "I lose balance" [19]. No other instances of determiner omission have been found.

Table 1: *character information.*

Character	Movie	Year	Gender	Protagonist/antagonist	Comic relief	L1 of actor
Alexei	*Black Widow*	2021	Male	Protagonist	Yes	English
Antonov	*Stranger Things*	2022	Male	Protagonist	No	German
Dreykov	*Black Widow*	2021	Male	Antagonist	No	English
Illya	*The Man from U.N.C.L.E.*	2015	Male	Protagonist	Yes	English
Melina	*Black Widow*	2021	Female	Protagonist	No	English
Natasha	*Black Widow*	2021	Female	Protagonist	No	English
Obruchev	*No Time to Die*	2021	Male	Antagonist	Yes	Danish/ Swedish
Oleg	*The Man from U.N.C.L.E.*	2015	Male	Antagonist	No	Russian
Safin	*No Time to Die*	2021	Male	Antagonist	No	English
Yelena	*Black Widow*	2021	Female	Protagonist	No	English
Yuri	*Stranger Things*	2022	Male	Antagonist	Yes	Serbian

Table 2: *article omissions.*

Character	Comic relief	Article omissions	Total articles	Percentage omitted
Alexei	Yes	12	66	18.2 %
Antonov	No	7	45	15.6 %
Dreykov	No	3	46	6.5 %
Illya	Yes	29	73	39.7 %
Melina	No	8	62	12.9 %
Natasha	No	0	59	0 %
Obruchev	Yes	7	20	35 %
Oleg	No	0	4	0 %
Safin	No	1	35	2.9 %
Yelena	No	3	82	3.7 %
Yuri	Yes	33	51	64.7 %

4.1. Statistics

The total number of articles follows a normal distribution as seen in Figure 1. The number of article omissions diverges from the normal quantile line when there are many article omissions, see Figure 2. The rest of the article omission data is randomly distributed around the normal quantile line. As such the data is assumed to follow a normal distribution. It is important to note that here as well as in the following, it is the total number of article omissions that are being investigated, *not* the percentage of omitted articles out of the total number of articles. For both the number of articles and the number of article omissions, the sample size is n=11.

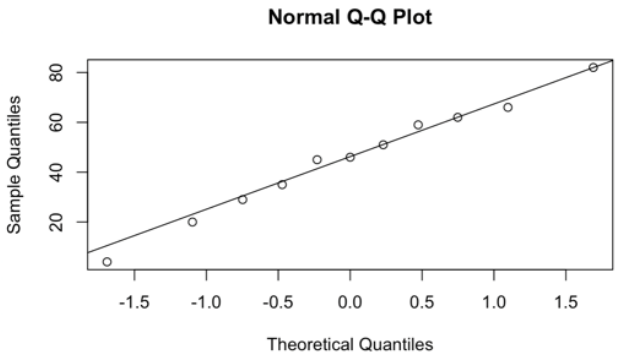


Figure 1: *Q-Q plot for articles.*

© 2024 Freja Hovgaard Knudsen. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

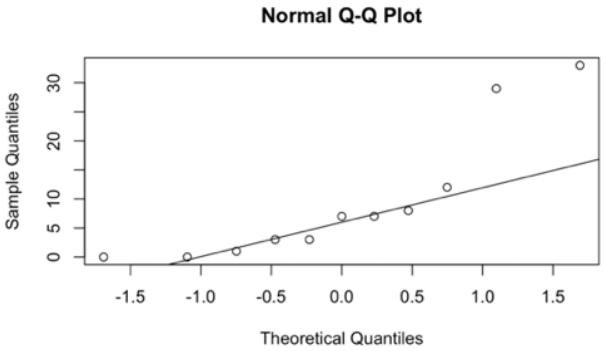


Figure 2: *Q-Q plot for article omissions.*

In R, vectors have been created for Omissions, Articles (combined total of present and omitted articles), and Year of movie release. The vectors with information about Gender, Protagonist vs Antagonist, Movie, Comic Relief, and L1 (simplified into 'English' and 'Other') have been made into factors.

Table 2 provides an overview of the articles, the article omissions, and the percentage of article omissions in relation to whether the character is a comic relief character or not. The data in the table seem to suggest a correlation between the number of article omissions and the percentage of article omissions and being a comic relief character. However, if untested we cannot 1.) claim this correlation as significant and 2.) dismiss the possibility that the correlation is between the number of article omissions and one of the other variables instead. To test which variables influence the number of article omissions, the model in which Omissions depend on Articles, Year, Comic Relief, Protagonist vs Antagonist, Movie, Gender, and L1 has been reduced to the model in which Omissions depend only on Comic Relief through the process of backwards selection. An F-test obtained through ANOVA in R does not oppose this reduction (p=0.2293). The amount of article omissions is positively correlated with being a comic relief character: Comic relief characters omit more articles than characters who are not categorised as comic relief characters. Figure 3 shows box plots of the number of article omissions for non-comic relief characters and for comic relief characters.

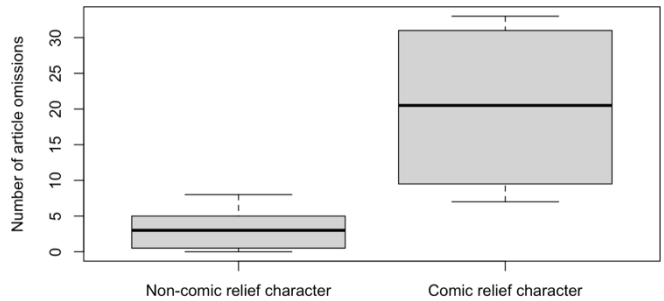


Figure 3: *boxplot of article omissions sorted by comic relief.*

Similarly, the model in which Articles depend on Year, Comic Relief, Protagonist vs Antagonist, Movie, Gender, and L1 has been reduced to the model in which Articles depend on Year and Protagonist vs Antagonist. An F-test obtained through ANOVA in R does not oppose this reduction (p=0.3675). Articles are positively correlated with Year and with being a Protagonist. The later the movie has been released, the greater the total amount of articles the character produces (cf. Figure 4), and a protagonist produces more articles than an antagonist (cf. Figure 5).

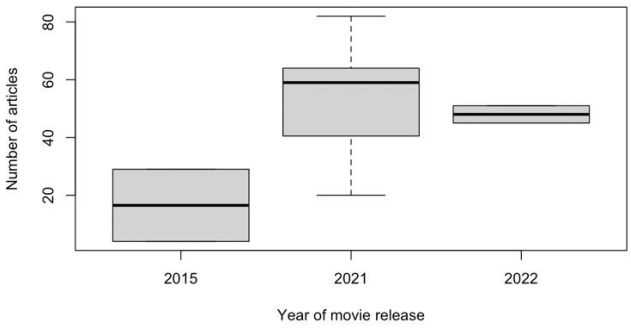


Figure 4: *boxplot of articles sorted by year of release.*

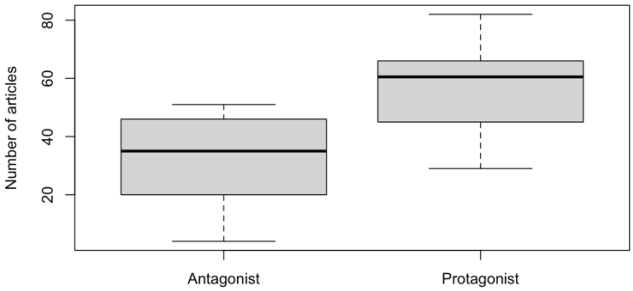


Figure 5: *boxplot of articles sorted by Protagonist vs Antagonist.*

Potential confounds might exist between the two variables Year and Protagonist vs Antagonist and a third variable not accounted for, namely the number of total words. It might be that characters in recent movies, in general, produce more words, or that protagonists say more than antagonists. In both cases, it might be that a greater number of total words contribute to a greater number of articles. Including this variable in further studies could be beneficial.

The reduction of the two models with Omissions only depending on Idiocy and Articles not depending on Idiocy, surprisingly indicates that the amount of article omissions does not rely on the total number of articles.

5. Discussion

The quantitative data strongly suggest a correlation between the amount of article omissions and idiocy. A more qualitative approach unveils the same tendencies.

In *Stranger Things* [20], comedic idiot Yuri is not introduced until halfway through episode 4, quickly establishing himself as an unreliable jokester through his horrendous histories meant as jokes, rapidly changing voice quality, and wheezing laugh. After his introduction, the comic relief of the series' plotline set in Russia relies heavily on him and the American hero Murray. Before Yuri's introduction, Antonov, while still not fully a comic relief character, contributed more significantly to the comic relief of that storyline: He is more jokingly antagonistic towards

Hopper, betting on his odds of escape and generally seeming almost carefree, a contrast to Hopper's desperation to escape. It is during this time that 5 out of Antonov's 7 article omissions occur. Furthermore, it is also at this point Antonov's two determiner omissions happen. After Yuri's debut, Antonov is imprisoned alongside Hopper, he becomes less upbeat, and is, as a typical sidekick, largely used to prompt Hopper to speak about his plans and feelings.

Throughout the series, Yuri continuously provides comic relief, but further characterisation and depth are never explored. Only in the very last episode, seriousness is added to his character when Antonov discloses Yuri's backstory in a conversation between the two. The conversation happens in Russian. Afterwards, Yuri does not have any dialogue. Importantly, this means that Yuri does not have any article omissions in his only emotionally potent conversation nor temporally close following it. This could be an indication that Yuri's way of speaking and his amount of article omissions are not fitting for a serious hero-like character.

In *No Time to Die* [21], Obruchev generally omits articles. However, one scene differs: Obruchev is interrogated by Bond and asked to explain how a weapon works. During this scene, Obruchev omits no articles out of 5. Here, Obruchev is more in control than he is in the rest of the movie – there is no physical fight, and he is talking about science, a topic in which he is knowledgeable. Providing important plot information, Obruchev's role as a comedic idiot is suspended. Perhaps, for this reason, article omissions are forfeited. Another explanation is an attempt to minimize the risk of the audience not understanding the information provided by Obruchev.

In *Black Widow* [19], Alexei is first reintroduced as an inmate in a prison. While imprisoned, he does provide some comic relief, but he is arguably too violent and fits his surroundings too well to fully be classified as a comedic idiot. While he is in prison, Alexei omits no articles out of 8. In the family reunion scene, Alexei fully embodies the comedic idiot persona and omits 6 out of 16 articles.

6. Conclusion

The quantitative article omissions data as well as the qualitative observations support the hypothesis that article omissions index humour and ridicule. A character that does not need to be taken too seriously and is meant to trigger amusement omits more articles than competent and emotionally important characters. This finding conforms with studies of Russian characters in animated films that have found that 'exaggerated and contrived Russian-accented English' is used to 'provide humour and to contrast with other characters' [7, p. 199].

For further studies, it would be worth examining if other features typical of Russian English are also distributed unequally among different types of characters thus indicating underlying indexes. Lawless's (2014) findings of used features in *Bond* films from 1962-2012 provide a great selection as these features are already proven to be present in at least some telecinematic productions [6, p. 88].

One of Lawless's observations is the 'use of sound [v] instead of [w]' [6, p. 88]. As this feature is also typical for German English [24] and both Russians and Germans have traditionally served as villains in telecinema [2; 8; 25; 26], an investigation into whether /w/ as [v] correlates with the character being an antagonist would be interesting. A further indication that it might be so, is that /w/ as [v] is also typical of Danish English [27] and while Danish characters are generally rare, recent years' telecinematic productions have seen a rise in the use of Danish actors for the antagonists [28; 29].

Regarding the eleven characters examined in this study, non-standard /h/ sounds have been detected for several characters, especially male characters, suggesting it might be worth looking into the distribution of this linguistic feature to investigate if it has indexes beyond that of Russianness. The effect of intonation patterns is also a possible research avenue to explore.

Generally, close attention to the linguistic variation in films strongly suggests that the variation depends on more than just the vernacular and L1 of the actor or the supposed factual background of the character as variation can be observed that is not easily explained by either of these things. This is also true of variation not related to the L1 of the characters as seen e.g., in the portrayal of Natasha and Yelena in *Black Widow* [19].

Yelena, who is introduced to the Marvel Cinematic Universe in *Black Widow*, has a hoarser less traditionally feminine voice quality than Natasha. Her clothes are baggier, and her hair is pinned up in a childlike braid. Natasha on the other hand has a softer voice quality, tight-fitting clothes, and loose hair. Unlike Yelena, Natasha was established within the Marvel Cinematic Universe prior to *Black Widow* and has on multiple occasions been cast as the love interest for various male heroes and has been seen using her gender and sexuality as a weapon, thus resembling the classic femme fatale figure. This is not the case for Yelena who is not seen to engage in any romantic or sexual behaviour. For both characters, their visual appearance and linguistic productions help aid characterisation and prepare the audience for what sort of behaviour can be expected of them in the storyline. In this manner, the indexicality of various linguistic features is used to create characters, manage expectations, and reinforce character tropes and genre conventions. It is therefore important to examine which features are used by which types of characters as this allows some insight into what the indexes have been at a given point and how they may be changing.

7. Acknowledgments

A big thank you to the Aarhus International Voice Studies Conference team, and especially to Míša Hejná for her supervision in the early stages of this research.

8. References

1. Bednarek, M. (2023). *Language and Characterisation in Television Series: A Corpus-Informed Approach to the Construction of Social Identity in the Media*. John Benjamins Publishing Company.

© 2024 Freja Hovgaard Knudsen. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Bleichenbacher, L. (2008). *Multilingualism in the Movies*. Göttingen: Hubert & Co.
3. Piazza, R., Bednarek M., & Rossi, F. (2011). Chapter 1: Introduction: Analysing telecinematic discourse. In Piazza, R., Bednarek M., & Rossi, F. (eds.), *Telecinematic Discourse: Approaches to the language of films and television series* (pp. 1-17). John Benjamins Publishing Company.
4. Colantoni, L., Steele, J., & Escudero, P. (2015). *Second language speech.* Cambridge University Press.
5. Lippi-Green, R. (2011). *English with an accent: Language ideology and discrimination in the United States*. Routledge. doi:10.4324/9780203348802
6. Lawless, K. (2014). 'Constructing the "Other": Construction of Russian Identity in the Discourse of James Bond Films'. *Journal of Multicultural Discourse*, 9(2), 79-97. doi:10.1080/17447143.2014.894517
7. Minutella, V. (2021). *(Re)Creating Language Identities in Animated Films*. Palgrave Studies in Translation and Interpreting. doi:10.1007/978-3-030- 56638- 8_4
8. Reenberg, N. (2022, March 2). *Stærk accent og koldblodige planer. I film er russerne altid skurke* [Strong accent and cold blooded plans. In movies, the Russians are always villains]. Kristeligt Dagblad. www.kristeligt-dagblad.dk/kultur/staerk-accent-og-koldblodige-planer-i-film-er-russerne-altid-skurke
9. Zavyalova, V., Proshina Z., Ionina A., Eddy A., & Ivankova, T. (2016). 4: Linguistic Features of Russian English. In Proshina Z. G. & A. A. Eddy (eds.), *Russian English: History, Functions, and Features*, (pp. 81–120). Studies in English Language. Cambridge: Cambridge University Press. 10.1017/CBO9781139683623.005.
10. Salinas, G. (2022, December 29). *Comic relief: Its definition and why it's necessary*. Videomaker. www.videomaker.com/how-to/planning/writing/comic-relief-its-definition-and-why-its-necessary/?amp=1
11. Werner, V. (2022). Pop Cultural Linguistics. *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. doi:10.1093/acrefore/9780199384655.013.999
12. Bell, A. & Gibson, A. (2011). Staging Language: An Introduction to the Sociolinguistics of Performance. *Journal of Sociolinguistics*, 15(5), 555-572. doi:10.1111/j.1467-9841.2011.00517.x.
13. Hodson, J. (2014). *Dialect in Film & Literature.* Palgrave Macmillan.
14. Silverstein, M. (2003). Indexical Order and the Dialectics of Sociolinguistic Life. *Language and Communication*, 23(3-4), 193-229. doi:10.1016/S0271-5309(03)00013-2
15. LeBlanc, R. J. (2021). Doing voices: Stylization, literary interpretation, and indexical valence. *Linguistics and Education,* 64, 1-11. doi:10.1016/j.linged.2021.200949
16. Coupland, N. (2001). Dialect Stylization in Radio Talk. *Language in Society*, 30(3), 345-375. doi:10.1017/S0047404501003013
17. Statista research department. (2023, January 5). *Most popular movie genres among adults in the United States as of December 2018, by gender.* Statista. www.statista.com/statistics/254115/favorite-movie-genres-in-the-us/
18. Kozloff, S. (2000). *Overhearing Film Dialogue.* Berkeley: University of California Press.
19. Shortland, C. (Director). (2021). *Black Widow* [Film]. Marvel Studios.
20. The Duffer Brothers (Producers). (2016 – present) *Stranger Things* [TV series]. Netflix.
21. Fukunaga, C. (Director). (2021). *No Time to Die* [Film]. Metro-Goldwyn-Mayer and Eon Productions.
22. Ritchie, G. (Director). (2015). *The Man from U.N.C.L.E.* [Film]. Warner Bros. Pictures and Ratpac Entertainment.
23. Rossi, F. (2011). Chapter 2: Discourse analysis of film dialogues: Italien comedy between linguistic realism and pragmatic non-realism. In Piazza, R., Bednarek M., & Rossi, F. (eds.), *Telecinematic Discourse: Approaches to the language of films and television series* (pp. 21-46). John Benjamins Publishing Company.
24. Bohn, O.-S. & Best, C. T. (2012). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. *Journal of Phonetics*, 40(1), 109-128. doi:10.1016/j.wocn.2011.08.002
25. Attanasio, P. (1985, May 5). *The Reich Stuff: Movie Villains of the Century.* Washington Post. www.washingtonpost.com/archive/lifestyle/style/1985/05/05/the-reich-stuff-movie-villains-of-the-century/37627b2e-ff98-4e1d-bfc2-d4d9badc126b/
26. Kasnik, B. (2018, May 4). *Nazisten er også en sjov skurk* [The Nazi is also a funny villain]. Kristeligt Dagblad. www.kristeligt-dagblad.dk/bagsiden/nazisten-er-ogsaa-en-sjov-skurk
27. Bohn, O.-S. & Caudery, T. (2017). 11: English with a foreign accent. In *The Sounds of English*. Aarhus: Department of English, Aarhus University.
28. Mikkelsen, M. (2022, August 22). *Pilou Asbæk: Vejen til Hollywood er brolagt med skurkeroller* [Pilou Asbæk: The road to Hollywood is paved with roles as villains]. Kristeligt Dagblad. www.kristeligt-dagblad.dk/kultur/pilou-asbaek-vejen-til-hollywood-er-brolagt-med-skurkeroller
29. Jimmy Kimmel Live. (2021, March 3). *Mads Mikkelsen on Being Rihanna's Bitch & Always Playing the Villain* [Video]. Youtube. www.youtube.com/watch?v=BDJFI_YJf4E

DOI: 10.2478/9788366675513-008

Voice quality changes, issue arises: a new typological profile of phonation in Zhangzhou

Yishan Huang[1], *Michael Proctor*[2]

[1]University of Sydney, Australia
[2]Macquarie University, Australia

yishan.huang@sydney.edu.au, michael.proctor@mq.edu.au

ABSTRACT

Zhangzhou exhibits complex patterns of phonation whose interactions with tonal phonology are not yet fully understood. Four phonation types–breathy, creaky, modal, and falsetto voice–can be observed in the non-utterance-final context; their distribution is primarily determined by vowel quality but further constrained by pitch contour, syllable coda type, and specific tonal category. In general, a breathy voice occurs on high vowels, regardless of the pitch contour. A creaky voice dominates low vowels, but only in falling-pitched tones. Modal phonation occurs in mid vowels and nonfalling-pitched low vowels. However, the voice contrasts are neutralized to falsetto in the tonal environment where syllables end in obstruent codas and occur in tone 6 (Yinru). Such an encoding, though dynamic, presents regular patterns that are attestable using acoustic waveforms and spectral tilt differences of H1-H2, H2-H3, and H1-H3. The encoding of Zhangzhou phonation is different from those laryngeally-complex languages as found in Central America, and also from those registered languages as found in Southeast Asia. It thus has important implications to the cross-linguistic typology of phonation in the world's tonal languages, while raising a theoretical question as to how the Zhangzhou speech should be classified in terms of its phonation property.

Key words: phonation, encoding, typology, Zhangzhou

1. Introduction

Phonation, which involves periodic opening and closing cycles of the vocal cords, provides the primary source for all voiced speech sounds in human languages [1-4]. The opening phase occurs when the subglottal air pressure exceeds the supraglottal air pressure, whereby the elastic forces push the vocal cords apart and force air to escape from the glottis. Contrastively, the closing phase happens when the subglottal pressure drops to a certain level below the supraglottal pressure, the elastic and aerodynamic forces cause the vocal folds to snap shut, whereby the subglottal pressure can be built up for the next glottal cycle.

The vibrating rate and mode can be altered in accordance with changes in the length, tension, and mass of the vocal folds, muscle of the larynx, and other glottal adjustments, resulting in different phonation types that humans can produce and perceive [1-9]. For example, creaky phonation occurs when the arytenoid cartilages are drawn together, causing the vocal folds to be tightly compressed as an unusually short, thick, and slack mass and vibrating at an irregularly slow rate [1-5]. In contrast, a falsetto voice occurs when the vocal folds are stretched longitudinally and become particularly thin and small, resulting in increased longitudinal tension upon the vocal ligaments and raised pitch height, accompanied by the maximum length of the vocal cords [5-7]. In contrast, in breathy phonation, the cartilaginous end of the glottis is not entirely closed but has a tiny opening; therefore, the expiratory airstream can constantly pass through, resulting in turbulent aperiodic noise that can be perceived throughout the glottal cycle [2-4].

As a speech phenomenon, phonation performs an important function in human languages, especially in tonal languages that use contrastive pitch to distinguish lexical meanings and/or encode different grammatical information [7-17]. For example, languages in Southeast Asia have conventionally been classified into purely tonal languages, register languages, and those between tonal and register languages, depending on the ways in which they use phonation and tone in their sound systems [9-15]. These register languages redundantly use pitch, voice quality,

© 2024 Huang et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

vowel quality, and durational difference to distinguish contrastive categories of words [9-15]. For example, in Mon, an Austroasiatic language spoken throughout Myanmar and Thailand, high-registered vowels are articulated with high pitch, shorter duration, and modal voice, whereas chest-registered vowels are produced with a lower pitch, longer duration, and breathy voice [12]. In addition, each tone in Burmese is characterized by a particular phonation, such as a low, modal-voiced tone; a high, breathy tone; a high, creaky tone; and a very high, tense-voiced tone [15].

In addition, languages in other parts of the world are also often discovered to employ phonation as an important factor in their tonal systems, but in a different way. Specifically, the Otomanguean languages are generally referred to as laryngeally complex languages in the literature because they use tone and phonation types independently in their phonology, whereby the number of tonal contrasts can be multiplied [3], [4], [8], [16], and [17]. For example, San Felipe Jalapa de Díaz (Jalapa) Mazatec, an Otomanguean language of the Popolocan branch contrasts three phonation types (breathy, modal, and laryngealised) and three-level tones (low, mid, and high) independently, resulting in a total of nine-way contrasts in its suprasegmental system [16]. Chichimec, an Otomanguean language of the Mixtecan branch, possesses two tones that differ in pitch levels but can occur in all three phonation types, resulting in a total of six contrasts [17].

Zhangzhou Southern Min, a Sinitic dialect spoken in the southern part of Fujian province in mainland China, employs phonation in its tonal system in a different way from those described in Southeast Asia and Otomanguean languages. Phonation does not play a phonemic function in this dialect, but four different phonation types of creaky, breathy, falsetto, and model voice can be observed in the empirical data. Similar to the Otomanguean languages, phonation in Zhangzhou can be independent of tone because the same tone can bear different phonation types. For example, the phonation of modal, breathy, and creaky can all occur in tone 4 in Zhangzhou, which has a mid-falling [41] pitch contour in citation and a high-falling [51] pitch contour in the phrase-initial context. However, the independence does not multiply the number of tonal contrasts in Zhangzhou. Similar to register languages in Southeast Asia, phonation can also be associated with a specific tone or vowel in Zhangzhou. For example, the breathy voice dominantly occurs in high vowels, whereas the falsetto voice occurs exclusively in tone 6 and in the non-utterance-final context. However, in Zhangzhou, the same tone and the same vowel can be realized with different phonations. For example, falsetto, creaky, and modal phonation can occur in low vowels, which cannot be attested to in register languages. Given the diversity and complexity of voice quality encoding, an important question can be raised as to how Zhangzhou speech should be classified as a laryngeally complex language, a register language, a combination of these two, or something else.

Driven by this intriguing phenomenon of phonation, this study is designed to comprehensively explore how phonation is encoded in Zhangzhou Southern Min, specifically in the non-utterance-final context. Two main issues are addressed: (1) How can various factors constrain the distribution of different phonation types in Zhangzhou? (2) How can dynamic encoding of phonation be quantified and visualised using scientifically objective patterns? Building upon the acoustic utterances from 21 native speakers of Zhangzhou Southern Min, this study will substantially advance and extend our knowledge of phonation type as an important language phenomenon in this Sinitic dialect, while offering new insights into how segments (e.g., vowel quality and syllable coda) and suprasegments (e.g., tonal pitch and tonal category) interact to shape a highly dynamic and diverse profile of phonation in the world's tonal languages. It will also contribute well-attested linguistic data to the cross-linguistic typology of phonation in tonal languages while enlightening the discussion on how this Southern Min variety should be classified from the perspective of phonation.

2. Research design

2.1. Zhangzhou city

Zhangzhou is a prefecture-level city situated in the southern Fujian province of Southern China, covering an area of approximately 12,600 square kilometres with a population of approximately 5.10 million [18-20]. The colloquial language spoken by native Zhangzhou speakers is Southern Min, referred to as Zhangzhou speech in this study, which is mutually intelligible with Southern Min varieties of Xiamen, Quanzhou, and Taiwan, but is entirely unintelligible to other Chinese dialects, such as Mandarin, Hakka, Wu, and Cantonese. This language is also well known as Hokkien by its colloquial pronunciation for its homeland Fujian and is widely spoken in many countries in Southeast Asia, such as Malaysia, Singapore, Indonesia, and the Philippines, because of the long-standing maritime trade and immigration from Fujian since the 16 century [21] and [22].

Certain regional variations can be observed in the sound system among the eleven administrative areas of Zhangzhou City [23] and [24]. For example, tone 7, corresponding to Yangru tone in terms of the middle Chinese tonal category, is documented as a short high contour [4] in Longhai, a mid-level contour [33] in Changtai, a low rising contour [13] in Dongshan and Zhao'an, and a convex contour of either [121] or [131] transcription elsewhere. Since this study is not intended to conduct a socio-phonetic investigation of Zhangzhou speech; it thus restricts the locality to the inner area of the Longwen and Xiangcheng districts, which are historically-socially-culturally-linguistically-geographically representative of Zhangzhou [18] and [19]. This selection can maximally limit the regional-related variation, while ensuring the derived acoustic patterns truly represent Zhangzhou speech as an independent language variety.

2.2. Research data

The data used in this study were obtained by the first author on a formal field trip in 2015 [18]. Twenty-one native speakers (9 males and 12 females) from the urban area of the Longwen and Xiangcheng districts participated in the data collection, with the average age of 56.5 for males and 50 for females. They were selected based on a set of criteria including intellectual curiosity, physical condition, age, birthplace, language environment at

home, occupation, education, and competence of other language(s) [18]. For example, they were all born and grew up in the urban areas of Zhangzhou, with Zhangzhou Southern Min as their mother tongue since their childhood. Their parents and spouses are also native Zhangzhou speakers. The selection criteria ensure the informants are representative of Zhangzhou Southern Min. Their utterances were recorded in an acoustically absorbent room at Zhangzhou Hotel. Tokens to be elicited were all written in simplified Chinese characters and presented via PowerPoint, with one token on one slide, which enabled speakers to produce required utterances with a balanced and well-controlled intensity and speech rate. All recordings were digitized at a sampling frequency of 44100 Hz in Praat [25]. A corpus of about 160 monosyllabic tokens and 588 disyllabic tokens were elicited from each speaker, upon which this study was built.

2.3. Acoustic processing

The acoustic processing of the obtained field data was conducted in Praat. Tonally relevant duration was identified incorporating all elements except syllable onsets [18]. Phonation-relevant duration was set at the last 10% portion of the tonally relevant duration, upon which the waveforms and spectral slices were extracted [26]. The amplitude values of the first three harmonics, labelled as H1, H2, and H3, were quantified to examine how they change in accordance with the increase of frequency of harmonics, and also to assess how they change in accordance with the change of vowel quality. Three spectral tilt differences of H1-H2, H2-H3, and H1-H3 were calculated to uncover the regularity and pattern for the production of different phonation types.

2.4. Acoustic normalization

The acoustic signals are highly variable because they carry both linguistic and extralinguistic information [27]-[30]. This study initiates a z-score normalization approach, which shared a similar principle as that for F0 normalization [27-31], and the Lobanov [32] approach for the normalization of formant frequency to reduce the indexical variances. The formula is expressed as in (1):

$$Ai = (Xi-m)/s \quad (1)$$

Xi is an observed amplitude value at a given harmonic (H1, H2, and H3) of a given vowel (high, mid, and low), while the parameters m and s, separately, stand for the raw mean amplitude value and the standard deviation estimated from all sampling amplitude values over all tokens from a given speaker. Ai is the derived normalized amplitude value.

3. Zhangzhou speech

3.1. Zhangzhou syllables

A template of C(G)V(X) can be generalized for Zhangzhou syllables in which onset (C) and nucleus (V) are compulsory while glide (G) and coda (X) are optional to occur [18-20]. Table 1 presents the segmental inventory for individual syllable components, which incorporates 15 onsets, 2 prevocalic glides, 13 nuclei and 8 codas. As seen, oral vowels, nasalized vowels and syllabic nasals can occur as nuclei, while nasal consonants, postvocalic glides, and obstruent consonants can function as coda. Among oral vowels, a three-way contrast of High (/i, u/), Mid (/e, ɵ/), and Low (/ɛ, ɐ, ɔ/) can be captured in terms of tongue height, which serves as an important factor motivating the realization and constraining the distribution of tonal phonation in this dialect. Figure 1 presents the acoustic distribution of the seven oral vowels in Zhangzhou [20] and [33]. Additionally, four main syllable types of CV, CGV, CVC, CGVC, in which tonal pitch is transcribed using Chao's [34] notation system, with number 5 representing the highest pitch level of the speaker's pitch range and number 1 the lowest. These serve as the foundation to explore how vowel quality and syllable type constrain the encoding of phonation in this Sinitic dialect.

Table 1: *phoneme inventory for Zhangzhou syllables.*

Component		Phoneme
Onset	C	p, p^h, ɓ, t, t^h, ɗ, k, k^h, ɠ, ts, ts^h, s, z, ɦ, ʔ
Glide	G	j, w
Nucleus	V	i, e, ɛ, ɐ, ɔ, ɵ, u, ĩ, ɛ̃, ɐ̃, ɔ̃, m̩, ŋ̩
Coda	X	j, w, m, n, ŋ, p, t, k

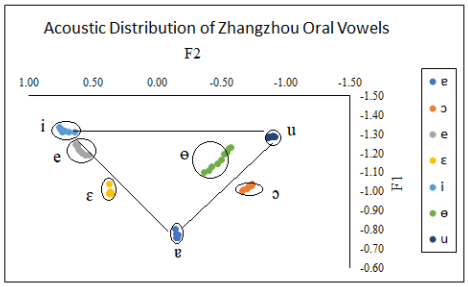


Figure 1: *acoustic distribution of Zhangzhou oral vowels.*

Table 2: *syllable types in Zhangzhou.*

Type	Example 1	Example 2	Example 3
CV	tĩ35 'sweet'	ʔɔ22 'lake'	ɓɐ41 'meat'
CGV	twɛ̃35 'bill'	kwɐ35 'song'	tshjɵ41'laugh'
CVC	sim35 'heart'	kut41 'bone'	kɐw41 'arrive'
CGVC	kwɐj41 'odd'	kjɐm22 'salty'	ʔjɐw35 'hungry'

3.2. Zhangzhou citation tones

Zhangzhou possesses eight lexical tones [18-20]. Contradictory to all prior impressionistically based studies that documented a seven-way tonal system [21-24] and [35-39], the eight-tone proposal is posited under two important phonetic facts. (a) Tones sharing a similar F0/pitch contour can differ considerably in other parameters, such as duration, vowel quality, syllable type and phonation. For example, tone 4 and tone 6 both have a mid-

© 2024 Huang et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

falling [41] pitch contour in citation, but the duration of tone 4 is statistically significantly shorter [18-19]. (b) Tones having an identical realization in citation can differ considerably in other linguistic contexts, such as at the non-right dominant (sandhi) position [18-19]. Figure 2 plots the normalised F0 pattern of the eight citation tones in Zhangzhou [18], [19], and [31], while Table 3 illustrates the eight citation tones with their pitch and duration values. The eight tones can be classified into stopped and unstopped categories depending on the syllable coda type that they are assigned to. Stopped tones (tones 6 and 7) can only be assigned to obstruent coda-ending syllables. All the other tones are assigned to sonorant-ending syllables, so that they are referred to as unstopped tones based on the Sinitic convention.

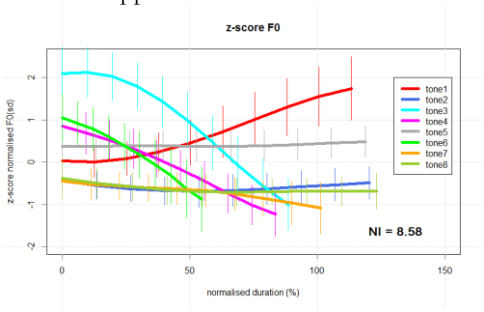


Figure 2: *normalized F0 patterns of Zhangzhou citation tones.*

Table 3: *examples of Zhangzhou citation tones.*

T	Pitch	Duration	Example 1	Example 2
1	[35]	extra-long	/tɐŋ/ 'east'	/kɔ/ 'aunty'
2	[22]	extra-long	/tɐŋ/ 'copper'	/kɔ/ 'glue'
3	[51]	medium	/tɐŋ/ 'to wait'	/kɔ/ 'drum'
4	[41]	medium	/tɐŋ/ 'frozen'	/kɔ/ 'oversee'
5	[33]	extra-long	/tɐŋ/ 'heavy'	/ɦɔ/ 'rain'
6	[41]	short	/tɐp/ 'answer'	/kɔk/ 'country'
7	[221]	long	/tsɐp/ 'ten'	/tɔk/ 'poison'
8	[22]	extra-long	/tsi/ 'tongue'	/kɔ̃ / 'snore'

3.3. Zhangzhou tone sandhi

Zhangzhou possesses a right-dominant tone sandhi system [18], [19], and [40]. The rightmost boundary of the sandhi domain is aligned with the rightmost boundary of a syntactic phrase XP, where X is categorically variable, ranging over adjective (A), noun (N), verb (V), preposition (PP), and adverbial (Adv) among others. Tones at the non-right-most (sandhi) position change their realisation phonetically and phonologically, while at the rightmost position maintain the realisation categorically similar to their citation forms but may be subject to certain variation at the phonetic level, because of the co-articulation and position effects. Figure 3 plots the normalised F0 patterns of individual sandhi tones as a function of their following tones based on the acoustic data from 21 speakers [40].

As seen, all tones change their realisation at the phrase-initial context. For example, tone 1 changes the F0 from a rising in citation to a mid-level contour in sandhi, while tone 3 changes to a rising from a high falling contour in citation. However, the F0/pitch realisation of the sandhi tones appears to be phonologically inert to its following tones because they present a similar tendency across tonal combinations. For example, tone 3 is consistently realised as a rising contour regardless of the F0/pitch contour of the tones that occur after.

Additionally, tones at the sandhi position can be neutralized severely [18] and [19]. For example, tones 1 and 2 neutralise their contours to be mid-level. Tones 5 and 8 neutralise the contrasts to be mid-falling. Tone 7 also shows a mid-falling tendency, but it has a significantly shorter duration. As well as this, the duration of sandhi tones is neutralized to two classes: those tones that are associated with sonorant codas neutralize their duration to be medium while the two stopped tones that are associated with obstruent codas neutralize their duration to be extra-short [18] and [19]. Thus, the F0 and duration system of the sandhi tones can be derived, as shown in Table 4, in which the tone level 6 is initiated by Huang [18] to denote the highest F0 level in the sandhi context, which is higher than the highest level that can be identified in the phrase-final context.

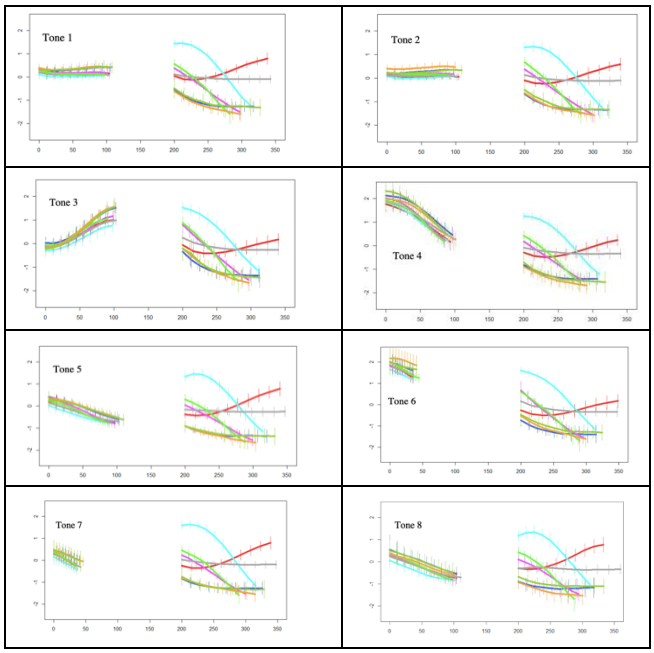


Figure 3: *normalized F0 patterns of sandhi tones across different tonal combinations.*

Table 4: *F0 and duration of Zhangzhou sandhi tones.*

T	pitch contour	F0	Duration
1	mid level	[33]	medium
2	mid level	[33]	medium
3	low rising	[24]/[25]	medium
4	extra-high-falling	[63]	medium
5	mid falling	[32]	medium
6	extra-high falling	[65]	extra-short
7	mid falling	[32]	extra-short
8	mid falling	[32]	medium

As indicated, the F0 inventory of the sandhi tones in Zhangzhou Southern Min incorporates five contour shapes of mid-level (tones 1 and 2), rising (tone 3), mid-falling (tones 5, 7, and 8), extra-high falling (tone 4), and extra-high and extra-short contour (tone 6). These also serve as the foundation to explore how tonal F0/pitch contour contributes to the dynamic encoding of phonation in Zhangzhou Southern Min, and also how the tonal category and the linguistic context can constrain the distribution of different phonation in this Sinitic dialect.

4. Zhangzhou phonation

Four phonation types comprising breathy, creaky, modal, and falsetto voice can be observed among phrase-initial (non-utterance-final) morphemes. Their distribution is found to be primarily conditioned by vowel quality but is further constrained by the tonal pitch contour, syllable coda type, and specific tonal category. In general, the breathy voice occurs only on high vowels; the creaky voice occurs on low vowels in falling-pitched tones; and the modal voice dominates in mid vowels regardless of pitch contour shape, and on those low vowels in non-falling-pitched tones. The falsetto voice, regardless of vowel height, occurs in the tonal environment where syllables end in obstruent codas and occur in tone 6 that corresponds to Yinru tone in terms of the Middle Chinese tonal category. Table 5 summaries the general tendency of the encoding of different phonation types in the phrase-initial (non-utterance-final) context in Zhangzhou, in which the symbol * indicates that no example can be attested in the empirical data of this study.

Table 5: *the distribution of phonation in the phrase-initial context in Zhangzhou.*

T	Pitch	Coda	High V	Mid V	Low V
1	[33]	sonorant	breathy	modal	modal
2	[33]	sonorant	breathy	modal	modal
3	[35]	sonorant	breathy	modal	modal
4	[63]	sonorant	breathy	modal	creaky
5	[32]	sonorant	breathy	modal	creaky
6	[65]	obstruent	falsetto	*	falsetto
7	[32]	obstruent	breathy	*	creaky
8	[32]	sonorant	breathy	modal	creaky

4.1. High vowel and phonation

The high vowels are dominantly produced with a breathy voice, regardless of the pitch contour shape, among those tones that are associated with sonorant-ending syllables in the phrase-initial context. Figure 4 shows the waveforms (the left-panel) of four phrase-initial morphemes that contain the same vowel /i/ but have different tones at the underlying level, while the right-panel presents their corresponding spectral slices. As seen, the waveforms of the four phrase-initial morphemes with the high vowel /i/ are all superimposed by dense, random, and also rapid glottal fluctuations across time and across four different contour shapes of rising [35], level [33], mid-falling [32], and extra-high falling [63]. Such manifestation indicates that the vocal folds are not completely closed over the articulation, whereby the airstream can freely pass through the glottis and create continuous turbulence [1], [2], and [9]. This is a typical articulatory basis for the breathy phonation.

In addition, as shown in the right panel, the amplitude values of the first harmonics are all significantly higher than their subsequent harmonics across tokens of varying pitch/F0 contour shapes, indicating a negative relation between the amplitude and frequency of harmonics. This property can be further justified by the normalised spectral tilt patterns from the 21 speakers, as shown in Figure 5. The H2-H3 and H1-H3 are both steeply positive across F0/pitch contours, signifying a dramatic drop of the amplitude in the low-frequency area. This reflects an increased open quotient and a less abrupt glottal closing gesture, causing energy to be distributed dominantly in the fundamentals and to decrease in the higher frequency regions [41], [42], and [43], which is a typical characteristic of breathy phonation. The value of H1-H2 is slightly higher than the mid-point, suggesting that the amplitude drop is not that significant in the second harmonic. Incorporating the waveform manifestation, the acoustic pattern of spectral tilt differences, it is plausible to propose that the high vowel is characteristic of a breathy voice in non-stopped tonal environments.

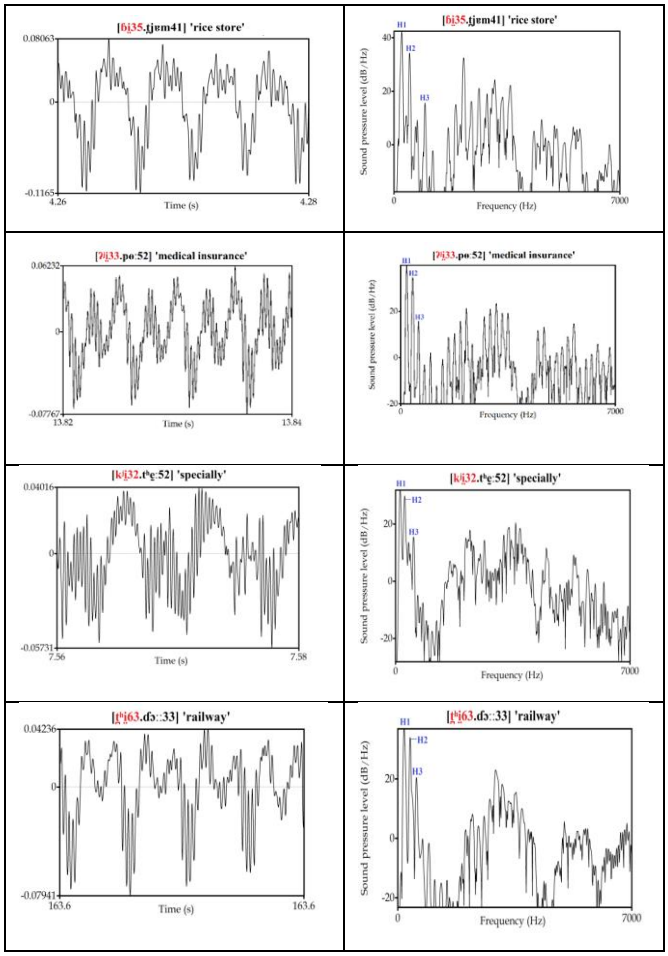


Figure 4: *waveforms and spectral slices of /i/ (WYF, male).*

© 2024 Huang et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

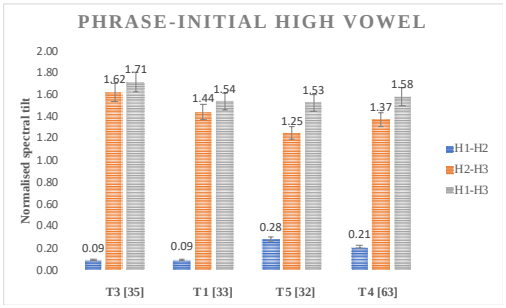


Figure 5: *normalized spectral tilt patterns of high vowel /i/.*

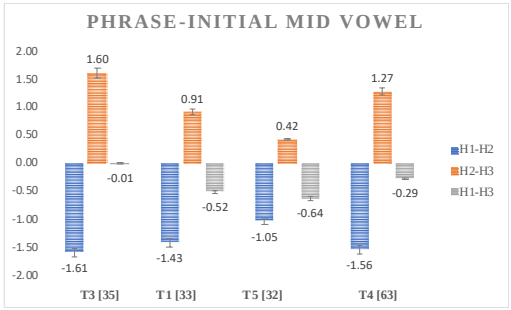


Figure 7: *normalized spectral tilt patterns of mid vowel /e/.*

4.2. Mid vowel and phonation

The phonation of mid vowels in the non-utterance-final context is observed to be modal to most speakers across different F0/pitch contours, but it may be articulated with a creaky voice by some male speakers in the falling pitched tones. The encoding of modal voice also has a robust basis in acoustics. As seen in Figure 6, the individual waveforms of the four phrase-initial morphemes containing the mid vowel /e/ are not superimposed by intense, random, and rapid friction across different F0 contours, denying a breathy phonation as found in the high vowel. Likewise, except in the third morpheme [te32], the intervals among glottal pulses are relatively evenly spaced, suggesting that the tokens are largely not articulated with a creaky voice by this male speaker WYF. Thus, it is proper to propose the mid vowel is related to modal phonation in Zhangzhou.

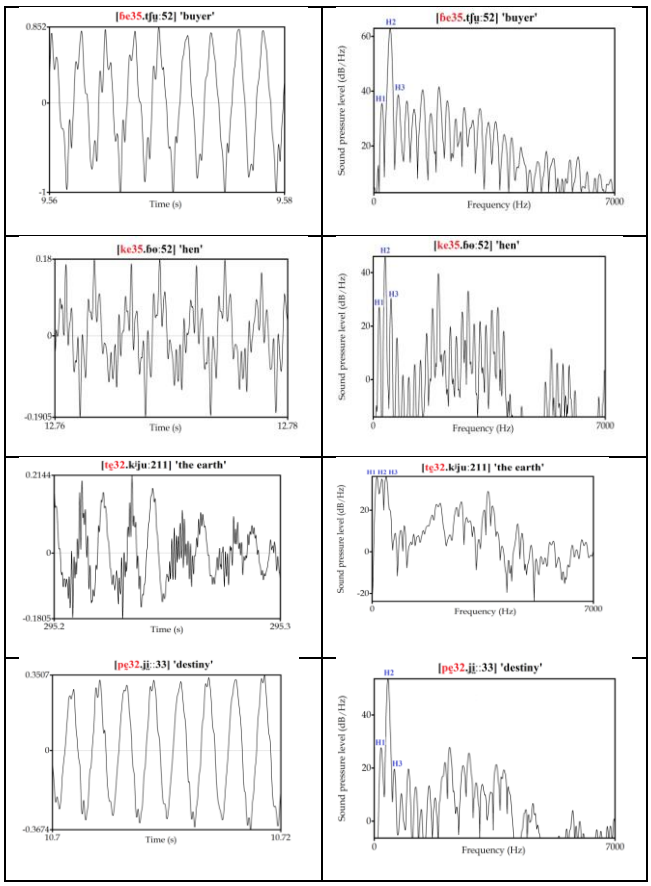


Figure 6: *waveforms and spectral slices of /e/ (WYF, male).*

In addition, as shown in the right panel of Figure 6, the amplitude value of the second harmonics is higher than that of both its preceding and following harmonics across most tonal pitch contours. This can also be robustly demonstrated by the normalized spectral tilt patterns that are quantified from the 21 native speakers' utterances. As shown in Figure 7, the H1-H2, and H1-H3 are steeply negative, while the H2-H3 are positive across the three different pitch/F0 contours. Thus, the dramatic increase of H2 amplitude and a dramatic drop of H3 amplitude constitute the central manifestation of the mid vowel with a modal-articulated phonation in the non-utterance-final context.

4.3. Low vowel and phonation

The high and mid vowels present only one dominant phonation type across those non-stopped tones in the phrase-initial environment. In contrast to this, the low vowels are found with voice quality alternating between modal and creaky phonations in this context. Their distribution is observed to be conditioned by the F0/pitch contour shapes of related tones. The phonation is shown to be modal when the related tones have a non-falling contour, while is creaky when the related tones are falling-pitched. This changing voice quality of low vowels can also be justified in acoustic patterns, as presented in Figure 8 and Figure 9.

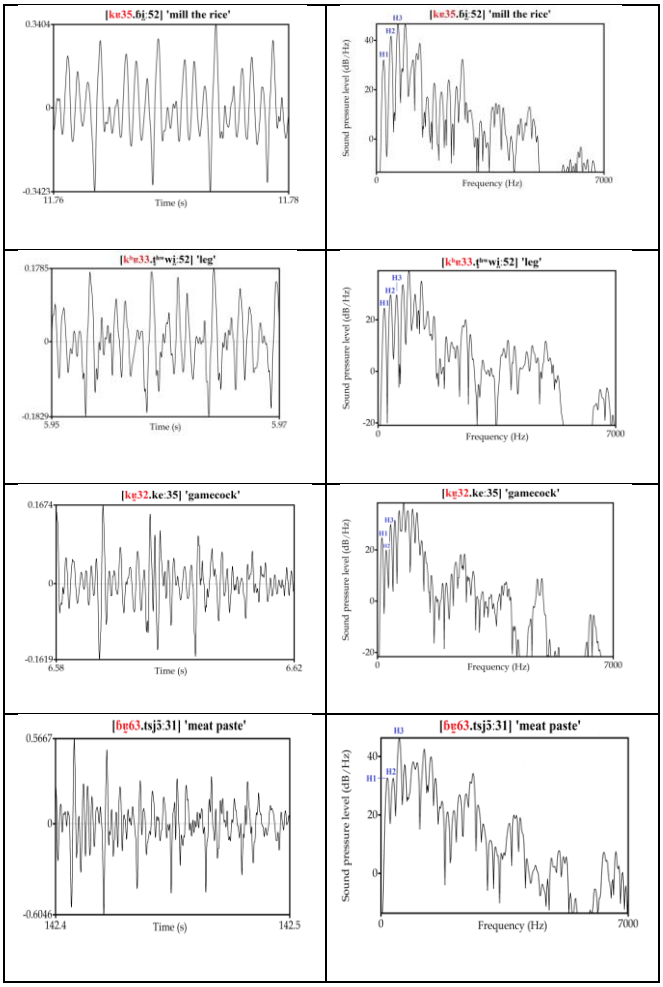


Figure 8: *waveforms and spectral slices of /ɐ/ (WYF, male).*

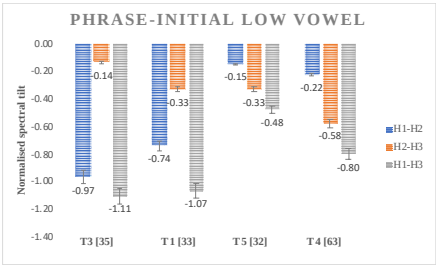


Figure 9: *normalized spectral tilt patterns of low vowel /ɐ/.*

In Figure 9, the normalized amplitude differences also vary considerably across tonal environments. (1) In the non-falling tones, the amplitude differences of H1-H2 and H1-H3 are steeply negative between 0.7 SD and 1 SD below the midpoint, while the H2-H3 difference is also negative, but slightly lower than the midpoint regardless of a rising [35] or level [33] pitch contour. This manifestation reflects an increasing tendency of the amplitude in accordance with the increases of frequency in the low frequency range, which is a typical characteristic of modal phonation. However, whether the H2-H3 should be positive or negative tends to vowel-dependent because it turns out to be positive for the mid vowel but negative for the low vowel in this phrase-initial context.

(2) In the mid falling [63] and the extra-high falling [32] contours, all the normalised spectral patterns of H1-H2, H2-H3, and H1-H3 are negative. The amplitude of the second harmonic appears very close to that of the first harmonic with the H1-H2 approaching zero (-0.2). On the contrary, the amplitude of the third harmonic is significantly higher than other lower harmonics, resulting in more deep negative values in H2-H3, and H1-H3. As indicated in Huang [18] and [25], a negative spectral tilt of H1-H2 near zero signifies a creaky phonation in this language, thus, it is plausible to assert that the low vowel is articulated creakily in the phrase-initial falling tones.

5. Phonation neutralisation

As discussed above, different voice quality of creaky, breathy, and modal are observable in the unstopped tones that are related to sonorant-ending syllables. Their distributions are primarily conditioned by vowel quality, in particular tongue height. However, vowels, regardless of their articulation height, are found neutralizing their voice contrasts to a new phonation type of falsetto when their related syllables end in obstruent codas and occur in the stopped tone 6, corresponding to Yinru tone in terms of the Middle Chinese tonal category. In this tone, no mid vowel can be attested; only high and low vowels can be found. The emergence of falsetto voice in this specific tonal context also presents some acoustic manifestations that can be quantified from native speakers' utterances. As seen in Figure 10, the waveforms of both high and low vowels are neither superposed with rapid, random, and intense friction, falsifying a breathy voice, nor did they present irregularly distributed and aperiodic glottal pulses, denying a typical creaky voice. Instead, the waveforms seem similar to that of modal phonation, but some slight differences can be observed. For example, the individual harmonics in the spectrum are easier to be identified because the curves look rounder and smoother than the ones in the modal phonation. Likewise, the number of glottal pulses within a given time (e.g., 0.02 seconds) is much greater than that in other contexts, indicating a much faster vibration of the vocal folds than the articulatory setting for the modal phonation [6-8].

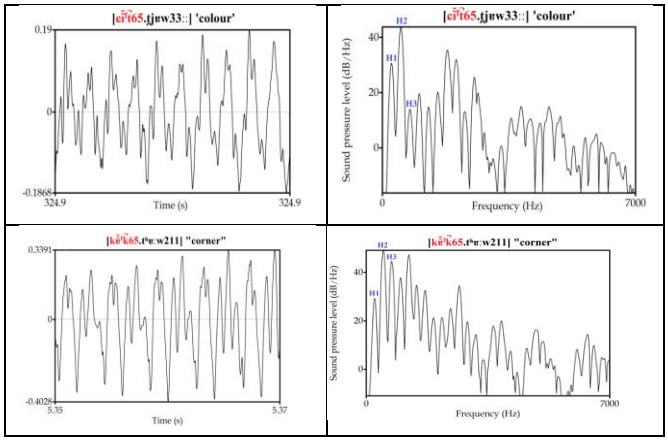


Figure 10: *waveforms and spectral slices of vowels in tone 6 (WYF, male).*

© 2024 Huang et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

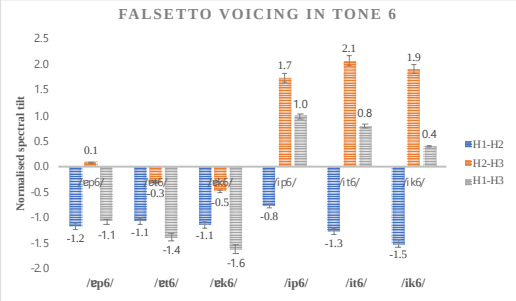


Figure 11: *spectral tilt pattern of vowels in tone 6.*

Additionally, the falsetto voicing in this phrase-initial tone 6 can also be justified by the normalised spectral tilt patterns. Figure 11 plots the 21 native speakers' phrase-initial utterances that contain high and low vowels across three different obstruent codas. As seen, the spectral tilt of H1-H2 is consistently negative across vowel height and obstruent coda type, but the H1-H3 and H2-H3 appear to be vowel-dependent. The low vowel presents a great negative value of H1-H3 below -1 SD and a negative H2-H3 value around the midpoint, regardless of the type of obstruent codas. On the other hand, the high vowel shows a steep positive value in H2-H3 above 1.5 SD, and a smaller positive H1-H3 within 0.3 and 1. Thus, though the spectral tilt patterns are shown to be dependent on vowel quality, they appear to be predictable and consistent within the same vowel and across syllables of different obstruent coda types. Thus, the articulation can be falsetto voicing and also be justified in acoustics.

The reason why vowels are articulated with a falsetto voice can be ascribed to the phonetics of its occurring environment. (a) The related syllables can only end in obstruent codas. The abrupt oral constriction for the obstruent coda articulation renders the related stopped tones to be realized with an extra short duration. This can be seen in Figure 12 that plots the normalised duration of the eight tones at the phrase-initial context [18-20] and [49]. Tone 5 has two duration bars because its realization is found to be sensitive to its following tones and presents two statistically significantly different variants. As seen, the two stopped tones 6 and 7 that are associated with obstruent-ending syllables are substantially shorter than any other tone.

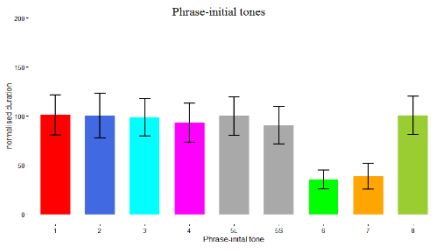


Figure 12: *normalized duration of phrase-initial tones.*

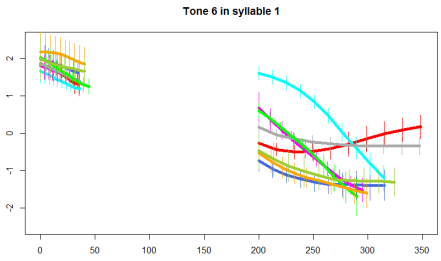


Figure 13: *normalized F0 contours of tone 6 across phrase-final tones.*

(b)The F0 contour of tone 6 in the phrase-initial context is shown to be extra-high. This can be seen in Figure 13 that plots its normalised F0 pattern across its eight following tones. The F0 range of tone 6 situates between about 1 SD and 2 SD, the height of which is much higher than the highest level that can be identified in the phrase-final setting. The extra-short-and extra-high F0 contour of this tone 6 provides an important articulatory base for the occurrence of the falsetto voicing, because the vocal folds in this case are stretched longitudinally, with the tension and length reaching their maximum upon the ligaments [18], [19], and [50].Therefore, incorporating the diverse phonetic factors of their waveform manifestations, spectral tilt patterns, the F0 contour and duration, and associated syllable coda type, it is plausible to propose a falsetto voicing, and phonation neutralization process in this special tonal environment.

6. Discussion

6.1. Phonation distribution

As discussed, the encoding of phonation in Zhangzhou is dynamic and diverse, with creaky, breathy, modal, and falsetto voice identifiable in the non-utterance-final context. Their occurrence is primarily conditioned by vowel quality but can be further constrained by the tonal pitch contour, syllable coda type, and specific tonal category. In general, the breathy voice exclusively occurs in high vowels, regardless of the pitch contour shape. The creaky phonation dominates in low vowels but only in the falling-pitched environment. The modal voice can be found in mid vowels regardless of the pitch contour, and in low vowels of non-falling pitch contours. However, vowels are observed neutralizing their voice contrast to falsetto in the special tonal environment where syllables end in obstruent codas and are assigned with tone 6.

While the encoding of phonation is severely constrained segmentally and suprasegmentally, their occurrence, in turn, can help enhance the perception and identification of particular segments and suprasegments. For example, if a segment is produced with a breathy voice, it is supposed to be high vowel in unstopped tones. If a segment is produced with a modal phonation, it is expected to be either a mid-vowel or a non-falling-pitched low vowel. Similarly, if a segment is produced with a creaky phonation, it is suggested to be low vowel in falling tones. As well as this, if a segment is articulated with a falsetto voice, this segment is supposed to be assigned to tone 6. In this

sense, phonation can also be considered reliable concomitant of tonal phonetics in Zhangzhou.

6.2. Phonation manifestation

The dynamic encoding of phonation that incorporates breathy, creaky, falsetto, and modal in Zhangzhou Southern Min is justifiable in acoustic and understandable from the articulatory perspective. Apart from their different waveform manifestations, the normalised spectral tilt patterns of H1-H2, H2-H3, and H1-H3 can well distinguish the existence of different phonation types in this Sinitic dialect. Based on the normalised spectral tilt patterns that are quantified from 21 native speakers, Table 6 summarises the encoding across different pitch/F0 contours and across different vowel height, in which *P* stands for a positive value for a particular spectral tilt difference, *N* for a negative value, and the letter *Z* for a tilt difference near zero. The symbol * signifies that no case can be attested in the empirical data.

Table 6: spectral pattern of H1-H2, H2-H3, and H1-H3 in the phrase-initial context.

Tone	Parameter	High V	Mid V	Low V
Tone 3 [35]	H1-H2	P/Z	N	N
	H2-H3	P	P	N/Z
	H1-H3	P	Z	N
Tone 1 [33]	H1-H2	P/Z	N	N
	H2-H3	P	P	N
	H1-H3	P	N	N
Tone 5 [32]	H1-H2	P/Z	N	N/Z
	H2-H3	P	P	N
	H1-H3	P	N	N
Tone 4 [63]	H1-H2	P/Z	N	N/Z
	H2-H3	P	P	N
	H1-H3	P	N	N
Tone 6 [65] Stopped	H1-H2	N	*	N
	H2-H3	P	*	N
	H1-H3	P	*	N

As indicated, the encoding of different phonation presents regular and predictable tendencies in terms of the normalised spectral title pattern. The breathy voice, which dominants on high vowels regardless of the pitch/F0 contours, presents a PPP (=positive H1-H2, positive H2-H3, positive H1-H3) pattern with a low H1-H2 value slightly above the midpoint at zero. The creaky voice, which occurs mainly on falling-pitched low vowels, shows a ZNN (=zero H1-H2, negative H2-H3, negative H1-H3) pattern with a salient feature of having H1-H2 somewhere just below the mid-point. The modal phonation has two variants. It presents an NPN (=negative H1-H2, positive H2-H3, and negative H1-H3) pattern for the mid vowels, but an NNN (=negative H1-H2, negative H2-H3, and negative H1-H3) pattern for those low vowels of non-falling pitched contours. The falsetto voice also presents two variants that are dependent on vowel quality. It presents a neat pattern of NPP (=negative H1-H2, positive H2-H3, positive H1-H3) for the high vowel but an NNN (negative H1-H2, negative H2-H3, and negative H1-H3) pattern for low vowel that can occur in tone 6. It seems that both the falsetto voice and the modal voice of low vowel share a similar NNN pattern, but they can be well distinguished in their waveforms as well as the associated pitch contour and duration. For example, the tone that is associated with the falsetto voice shows an extra-high and extra-short pitch contour. As a whole, the normalised spectral tilt patterns can well distinguish different realisations of phonation in this Southern Min variety. This offers a new methodological approach to quantify and visualise the phonation difference and reveal their relation with other segmental (vowel quality) and suprasegmental (tonal F0 and duration) parameters in tonal languages.

6.3. Phonation typology

Tonal languages in the world are generally classified into purely tonal languages, registered languages, and laryngeally complex languages in accordance with the function of phonation in related languages. The dynamic encoding of phonation in Zhangzhou Southern Min contributes a new profile to the typology of phonation in tonal languages. Similar to the Otomanguean languages, phonation in Zhangzhou can be independent of tones. For example, the breathy, creaky, and modal phonation can occur in the same falling tonal setting, such as in tone 4 and tone 5. In this sense, Zhangzhou could be regarded as a laryngeally-complex language. However, phonation does not perform a phonemic function and does not multiply the number of tonal contrasts in Zhangzhou. Similarly to the register languages in Southeast Asia, phonation in Zhangzhou can also occur in a specific tone, such as the falsetto voice can occur only in tone 6. In addition, phonation in Zhangzhou can also occur in a specific vowel. For example, the breathy voice can only be observed in the high vowels. In this sense, Zhangzhou could be considered as a registered language. However, unlike the register languages, the same tone in Zhangzhou can bear different voice qualities, for example, three different phonation types can occur in the same falling-pitched tones. Similarly, the same vowel can be realized with varying phonation. For example, the low vowel can be realised with three phonation types: a modal voice in non-falling pitch contour; a creaky voice in falling pitch contour, and a falsetto voice in tone 6.

Given the diverse and dynamic encoding, an important question may be raised as to how Zhangzhou speech should be classified: as a purely tonal language, a register language, a laryngeally complex language, a combination of these, or something else? This reflects the critical fact that languages can be encoded in a way that is far more complicated than our general assumptions and expectations. It is thus imperative to ground linguistic analysis in phonetic reality to uncover its nature and to improve our cognitive knowledge of human languages.

7. Conclusion

This study comprehensively explores how voice quality is encoded in Zhangzhou Southern Min, a Sinitic dialect spoken in South Fujian province of mainland China. Empirically, four phonation types of breathy, creaky, falsetto, and modal can be identified in the non-utterance-final context, whose distribution is primarily conditioned by the vowel height, and further constrained by pitch contour shape, syllable coda type, and a specific tonal category. The dynamic encoding is well attested in acoustic waveforms and normalised spectral tilt patterns of H1-H2, H2-H3, and H1-H3 from 21 native speakers' utterances.

© 2024 Huang et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

This study substantially advances and broadens our knowledge of how the segments and suprasegments interact to shape a dynamic profile of phonation in this Southern Min variety. It contributes a new methodology using the spectral tilt patterns to quantify and visualise the phonation differences in natural speech. This finding also sheds an important light on using other laboratory methods, like EGG [51] and GAMMs [52], to analyse the phonation variation in this Sinitic dialect. The dynamic encoding of phonation in Zhangzhou also offers a new typological profile of phonation in tonal languages, while highlighting a typological issue as to how this Southern Min dialect should be classified: as a register language as found in tonal languages in Southeast Asia, a laryngeally complex language as found in the Otomanguean family, a combination of these, or something else.

8. References

1. Laver, J. (1980). *Phonetic description of voice quality*. Cambridge, England: Cambridge University Press.
2. Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, 29, 383-406.
3. Esposito C., & Khan, S. (2020). The cross-linguistic patterns of phonation types. *Language and Linguistics Compass*, 14 (12), 1-25.
4. Esposito, C. M. (2012). An acoustic and electroglottographic study of White Hmong tone and phonation. *Journal of Phonetics*, 40, 466-476.
5. Akita, K. (2021). Phonation type matter in sound symbolism. *Cognitive Science*, 45(5): e129821.
6. Podesva, R. (2013). Gender and the social meaning of non-modal phonation. In Cathcart, C., Chen, I-H. Finley, G., Kang, S., Sandy, C. S., & E Stickles, *proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, 427-448.
7. Zhu, X. (2012). Multiregisters and four levels: A new tonal model. *Journal of Chinese Linguistic*s, 40(1), 1-17.
8. Esposito, C. M. (2006). *The effects of linguistic experience on the perception of phonation*. Doctoral dissertation: University of California, Los Angeles.
9. Andruski, J. E., & Ratliff, M. (2000). Phonation types in production of phonological tone: The case of Green Mong. *Journal of the International Phonetic Association*, 30, 37-61.
10. DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association*, 39, 162-188.
11. Kuang, J. (2013). *Phonation in tonal contrasts*. Doctoral dissertation: University of California: Los Angeles.
12. Brunelle, M., & Kirby, J. (2016). Tone and phonation in Southeast Asian languages. *Language and Linguistics Compass*, 10 (4), 191-207.
13. Brunelle, M., & Tấn, T. (2021). Register in languages of Mainland Southeast Asia: state of the art. In Sidwel, P., & M. Jenny (Ed.), *The Languages and Linguistics of Mainland Southeast Asia: A comprehensive guide.* (pp.683-706), Berlin, Boston: De Gruyter Mouton.
14. Brunelle, M. (2005). *Register in Eastern Cham: phonological, phonetic and sociolinguistic Approaches*. Doctoral thesis: Cornell University.
15. Mortensen, D.R. (2013). Tonally conditioned vowel raising in Shuijingping. *Journal of East Asian Linguistics*, 22, 189-216.
16. Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, 41, 185-205.
17. Kelterer, A., & Schuppler, B. (2020). Phonation type contrasts and tone in Chichimec. *The Journal of the Acoustical Society of America*, 147(4), 3043-3059.
18. Huang, Y. (2018). *Tones in Zhangzhou: pitch and beyond*. Doctoral thesis: Australian National University.
19. Huang, Y. (2020). *Tones in Zhangzhou: pitch and beyond. Cambridge*, UK: Cambridge Scholars Publishing.
20. Huang, Y. (2021). *Zhangzhou Southern Min: Syllables and Phonotactics*. Lincom Europa.
21. FJG. (1998). *Fujian Province Gazette-Dialect Volume. Beijing*, China: Fangzhi Chubanshe.
22. ZZG. (1999). *Zhangzhou City Gazette-Dialect Volume (Vol. 49)*. Beijing, China: Zhongguo Shehui Kexue Chubanshe.
23. Ma, C. (1994). *Studies of Zhangzhou dialect*. Hongkong: Zongheng Chubanshe.
24. Yang, X. (2008). *Studies of tones and regional cultures of Zhangzhou dialect*. Beijing, China: Zhongguo Shehui Kexue Chubanshe.
25. Boersma, P. (2001). Praat: Doing phonetics by computer [Computer program]. *Glot International*, 5, 341-345.
26. Huang, Y. (2023). Phonation Variation in Zhangzhou Southern Min. *Proceedings of 2023 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR 2023)*, September 22-24, Hangzhou, China.
27. Harrington, J. (2010). Acoustic phonetics. In Hardcastle, W., Laver, J. J., & F. E. Gibbon (eds.), *The handbook of phonetic sciences (2nd ed)* (pp.81-129). Hoboken, NJ: Wiley-Blackwell.
28. Huang, Y., Donohue, M., Sidwell, P., & Rose, P. (2016). Normalization of Zhangzhou citation tones. In Carignan, C., & M. Tyler (eds.), *Proceedings 16th Australasian International Conference on Speech Science & Technology*, 217-220. Sydney, The Australian Speech Science & Technology Association.
29. Rose, P. (1986). The normalization of tone. *Proceedings of the First Australian Conference on Speech Science and Technology*, Canberra, Australia, 130-135. Sydney: Australian Speech Science and Technology Association.
30. Rose, P. (2016). Comparing normalization strategies for citation tone F0 in three Chinese dialects. In Carignan, C. & M. Tyler (eds.), *Proceedings of the 16th Australasian International Conference on Speech Science and Technology*, 221-224. Sydney: Australian Speech Science and Technology Association.
31. Huang, Y. (2022). Pairwise T-Test: identification of tonal relations. *Proceedings of the International Conference on Asian Language Processing*, 27-28 October, Chinese University of Hong Kong (Shenzhen), China; the Chinese and Oriental Languages Information Processing Society, Singapore.
32. Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B), 606-608.
33. Huang, Y. (2023; forthcoming). Encoding Syllables in Zhangzhou Southern Min. *Proceeding of The 37th Pacific Asia*

Conference on Language, Information and Computation, 2-5 December, The Hong Kong Polytechnic University, China.
34. Chao, Y. (1930). A system of "tone letters". *Le Maître Phonétique*, 45, 24-27.
35. Dong, T. (1959). *Four Southern Min varieties*. Taipei: Zhongyang Yanjiuyuan.
36. Lin, B. (1992). Zhangzhou vocabularies. *Fangyan*, 1-3.
37. Zhou, C. (2006). *The great Southern Min dictionary*. Fuzhou, China: Fujian Renmin Chubanshe.
38. Guo, J. (2014). *Zhangzhou Southern Min*. Zhangzhou, China: Zhangzhou Library.
39. Gao, R. (1999). Introduction to the sound system of Zhangzhou. In *Minnan dialect-studies of Zhangzhou variety* , 109-116. Beijing, China: Zhongguo Wenlian Chubanshe.
40. Huang, Y. (2022). Right-Dominant tones in Zhangzhou: on and through phonetic surface. *Proceeding of the 34th annual Conference on Computational Linguistics and Speech Processing in Taiwan,* 21-22 Nov., Taipei Medical University, China.
41. Stevens, K. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica*, 34, 264-279.
42. Huffman, M. K. (1987). Measures of phonation type in Hmong. *Journal of the Acoustical Society of America*, 81, 495-504.
43. Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., & Goldman, S. L. (1995). Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech, Language, and Hearing Research*, 38, 1212-1223.
44. Johnson, K. (2011). *Acoustic and auditory phonetics (3rd ed)*. Hoboken, NJ: Wiley-Blackwell.
45. Laver, J. (1980). *Phonetic description of voice quality*. Cambridge, England: Cambridge University Press.
46. Lotto, A. J., Holt, L., & Kluender, K. R. (1997). Effect of voice quality on perceived height of English vowels. *Phonetica*, 54, 76-93.
47. Thurgood, G. (2000). Voice quality differences and the origin of diphthongs. *Annual Meeting of the Berkeley Linguistics Society*, 26, 295-303.
48. Baart, J. (2010). *A field manual of acoustic phonetics*. Dallas, TX: SIL International.
49. Huang, Y. (2022). Duration in Zhangzhou Southern Min: Variation, Correlation, and Constraint. *Proceedings of the 18th Australasian International Conference on Speech Science and Technology*. 13-16 December 2022, Canberra, The Australian Speech Science & Technology Association.
50. Huang, Y. (2023). Same F0, different tones: A multidimensional investigation of Zhangzhou tones. *Proceedings of INTERSPEECH 2023*, 4728-4732, 20-24 August, Dublin, Ireland.
51. Hong K. & Kim H. (1997). Electroglottography and laryngeal articulation in speech. *Folia Phoniatr Logop*. 149(5):225-33.
52. Carignan, C., Hoole, P., Kunay, E., Pouplier, M., Joseph, A., Voit, D., Frahm, J. & Harrington, J., (2020). Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI, *Laboratory Phonology* 11(1): 2.

© 2024 Huang et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-009

Voice your opinion! Does modality affect the acceptability of extraction from adjunct clauses in English?

Anne Mette Nyvad[1], *Ken Ramshøj Christensen*[1], and *Julie Maria Rohde*[1]

[1]Department of English, Aarhus University, Denmark

amn@cc.au.dk, krc@cc.au.dk, jmr@cc.au.dk

ABSTRACT

Even though adjunct clauses are traditionally assumed to be universal syntactic islands, extraction from finite adjunct clauses has been shown to be acceptable to varying degrees in the Mainland Scandinavian languages, as well as in English. The relative acceptability of extraction depends on several factors, including the type of adjunct clause, the type of extraction dependency, and presence of contextual facilitation. Based on previous studies on English and Danish, we conducted a series of acceptability judgment experiments (judgements on a 7-point Likert scale) on extraction from three types of finite adjunct clauses (corresponding to English *if*-, *when*- and *because*-clauses). Unlike previous studies, we wanted to explore whether people respond differently to spoken stimuli than to written stimuli. We used an AI with a male and a female voice to generate the stimulus sentences based on a previous study. We hypothesized that the spoken versions would be judged as more acceptable than their written counterparts, because such sentences are typically found in a colloquial spoken language. However, that is not what we found. Our results replicated a previous study with identical, but written stimuli, showing that the acceptability pattern associated with extractions from adjunct clauses appears to be stable across modalities.

Key words: participant gender, speaker voice, syntactic islands

1. Introduction

In a range of languages, including Danish and English, it is possible to find sentences in which an element that is interpreted inside the embedded clause has been extracted or 'moved' to the front of the clause, as in (1). As such, the extracted element can be said to lead a sort of 'double life':

(1) Which article did you say [that he had read __]?

However, in the literature on generative grammar, extraction from certain types of embedded clauses (e.g. relative clauses and indirect questions) is assumed to be ungrammatical across languages. In (2), a *wh*-element is extracted from an embedded question, a relative clause, and an adjunct clause, respectively:

(2) a. *How did John ask [who behaved __]? [1]

b. *Who does Phineas know a girl
[who is working with __]? [2]

c. *Who did Mary cry [after John hit __]? [3]

Ross [2] calls such constructions syntactic 'islands' as they are difficult or impossible to leave (cf. the story of *Robinson Crusoe*).

The focus of this paper is on extraction from adjunct clauses, cf. (2c). The ungrammaticality of such constructions has been assumed to be due to one or more universal constraints [3–5], and as such, we would expect low acceptability of sentences violating it, as well as little or no variation across constructions and languages. However, in previous experimental studies, we have found that extraction from *wh*-clauses [6,7] and from relative clauses [8] appears to be grammatical in Danish, contra generative theoretical syntax. Recent studies have revealed that extraction from adjunct clauses also appears to be possible in the Mainland Scandinavian languages, but that there is a great deal of variation in acceptability which depends on (i) the adjunct clause type (topicalization out of conditional and temporal clauses is generally better than extraction from a clause of reason [9,10]), (ii) the type of dependency (topicalization is better than *wh*-extraction in these languages [11,12]), and (iii) context (the addition of a facilitatory context appears to increase the acceptability of extraction out of adjunct clauses [12]). Based on these findings, Nyvad et al. [13] decided to investigate extraction from adjunct clauses in English, cf. (2c), to see if the effects found in the Mainland Scandinavian languages would also show up here. Instead of topicalization, we chose to employ relativization out of the adjunct clauses, given that a prior corpus study of English found naturally occurring examples of adjunct clause extractions by way of this dependency type only. The

© 2024 Anne Mette Nyvad et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

results of this acceptability study revealed a great deal of variation: Relativization out of conditional adjunct clauses introduced by *if* was not significantly less acceptable than extraction out of complement *that*-clauses. This finding is quite surprising because the latter type of clause is not an island, whereas the former is usually considered to be one. Extraction from *if*-clauses was significantly more acceptable than that from temporal *wh*-clauses, which in turn was significantly more acceptable than relativization out of a clause of reason introduced by *because*. All of the extractions from adjunct clauses were significantly more acceptable than extraction from a coordinate structure. Based on these findings, we concluded that adjunct clauses are not strong islands in English.

We present the results of a study employing the same stimuli as in Nyvad et al. [13], but exposing participants to auditory stimuli generated with AI instead of written stimuli, in order to see whether there is a difference in responses. Our hypothesis was that the spoken stimuli would be judged as even more acceptable than the written ones, given that island violations are assumed to be a primarily spoken language phenomenon (see [14,15] for Swedish, and [16] for Danish and English).

We had five research questions:

- **Q1**: What is the overall acceptability pattern?
- **Q2**: Is the overall acceptability pattern for the spoken stimuli different from the pattern for the written stimuli?
- **Q3**: Are there any (participant) gender or (speaker) voice effects on acceptability?
- **Q4**: What is the overall RT pattern?
- **Q5**: Any (participant) gender or (speaker) voice effects on RT?

2. Data and methods

We conducted an experiment using a subset (to make the experiment shorter) of the set of sentences as in Nyvad et al. [13], namely relativization out of complement clauses with *that* and adverbial clauses introduced by *if*, *when* and *because*, compare (3) and (4):

(3) a. It's obvious that I was surprised
[**that** she actually completed this exercise].

b. It's obvious that I would be surprised
[**if** she actually completed this exercise].

c. It's obvious that I was surprised
[**when** she actually completed this exercise].

d. It's obvious that I was surprised
[**because** she actually completed this exercise].

(4) a. This is the exercise that I was surprised
[**that** she actually completed __].

b. This is the exercise that I would be surprised
[**if** she actually completed __].

c. This is the exercise that I was surprised
[**when** she actually completed __].

d. This is the exercise that I was surprised
[**because** she actually completed __].

In order to explore the potential impact of differences between individual voices, we used a female and a male voice (though we had no a priori prediction about any effect of voice gender on acceptability or RT). To avoid potential human speaker bias, we used an AI to generate speech (https://elevenlabs.io/). On the website, the female voice that we used, *Bella*, is described with the terms 'soft' and 'narration', and the male voice, *Charlie*, is described as 'casual, conversational'. To us, both voices sound pleasant, casual, and calm.

There were 48 target sentences in total (plus 28 fillers not discussed here). We used a Latin square design with two lists (controlling for speaker voice, extraction, and type), such that each participant judged 38 items. Participants were pseudo-randomly assigned (based on their birth month) to one of the two lists. The experiment was set up in PsychoPy (v.2023.1.3) [17,18] and hosted online by Pavlovia.org. The design is shown in Figure 1. The task was a binary acceptability judgment task, where participants had to judge whether each individual sentence was 'Not OK = unacceptable' or 'OK = acceptable'. The participants were instructed to respond as fast and as accurately as possible and to base their judgments on their own intuitions, not on what they might expect to be 'correct' or 'standard' language. Before the actual experiment they received a training session with four sentences to familiarize themselves with the task. The option to respond only became possible after each individual sentence had been spoken.

A total of 48 native speakers of English (24 female gender, 24 male) volunteered to participate in the experiment. The age of the participants ranged from 17 to 72 years (mean = 44.2, SD = 15.8). 19 participants evaluated the items on list 1, and 29 evaluated the items on list 2. The data was analyzed using *R* version 4.1.2 [19] with the *lmerTest Package* [20] and the *lme4* package [21] for mixed effects models of continuous and binomial data, respectively, and the *MASS Package* [22] for sliding, pair-wise contrasts. All plots were made using the *ggplot2* package [23]. Prior to analysis, we removed data points (31 of the total 1,824) with response times ≥6 seconds (too slow for speeded acceptability).

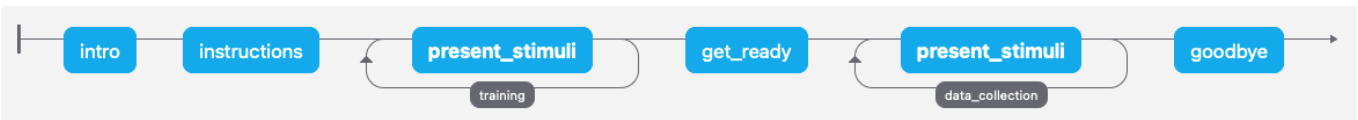


Figure 1: *experimental design in PsychoPy.*

3. Results: acceptability

3.1. The overall pattern for acceptability

Figure 2 shows the overall mean acceptability rating for each clause type with and without extraction. (The mean rating for filler 1 was 0.96, and 0.04 and 0.03 for filler 2 and filler 3, respectively.) The data was fitted to a mixed-effects model with the binomial answer as outcome variable and type as predictor (with sliding contrasts) and random intercepts for participant and item, and random slopes for type by participant and type and trial by item. The summary of the results is provided in Table 1.

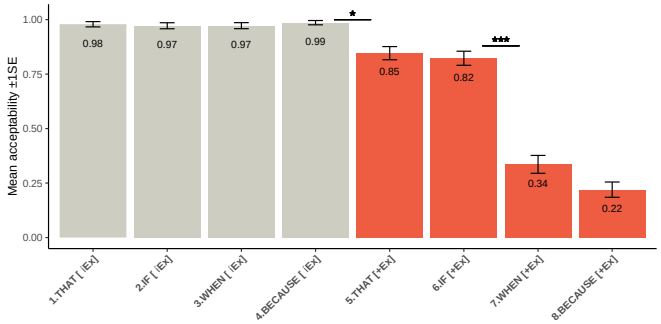


Figure 2: *acceptability rating by clause type. Sliding contrasts; *** p < 0.001, * p < 0.05. The binary acceptability rating is averaged across items and participants.*

Table 1: *summary of the mixed-effects model for acceptability with sliding contrast. *** p <0.001, * p< 0.05.*

	Estimate	SE	z	p
1. *that* [–Ex] vs. 2. *if* [–Ex]	-4.329	2.41	-1.80	0.072
2. *if* [–Ex] vs. 3. *when* [–Ex]	1.714	1.57	1.09	0.276
3. *when* [–Ex] vs. 4. *because* [–Ex]	0.226	1.87	0.12	0.904
4. *because* [–Ex] vs. 5. *that* [+Ex]	-4.165	1.63	-2.56	0.011
5. *that* [+Ex] vs. 6. *if* [+Ex]	0.773	1.07	0.72	0.472
6. *if* [+Ex] vs. 7. *when* [+Ex]	-6.978	1.25	-5.60	0.000
7. *when* [+Ex] vs. 8. *because* [+Ex]	-0.095	1.27	-0.08	0.940

3.2. Modality effect on acceptability

To test for a potential effect of spoken vs. written modality, we compared the z-transformed ratings from the current study to z-transformed data from Nyvad et al. [13]. A participant's z-transformed rating represents the number of standard deviations which the raw (non-transformed) rating is from that participant's mean rating. Z-transformation has two benefits: It controls for potential individual scale bias (e.g. using only one or both extremes on a scale), and it makes direct comparison between studies that may not use the same scale easier [24,25] – as is the case for the present comparison of responses on a binary scale to response on a 7-point Likert scale.

We subjected the data to a mixed-effects model using the *lmerTest Package* for *R* [20]. The model has type (1–8, cf. Figure 2), modality (spoken vs. written), and the interaction between the two as predictors. The model also had random intercepts for participant and item, random slopes for type and modality by participants, and random slopes for modality for items. As shown on Table 2, the interaction between type and modality was only significant for type 7, *when* [+Ex] ($p < 0.04$) and marginal for type 8, *because* [+Ex] ($p < 0.07$): Both were rated higher in the written modality. (The fixed 'main' effect for modality was marginal, $p < 0.08$.)

Table 2: *summary of the mixed-effects model for modality (Mod.).*

	Est.	SE	df	t	p
Modality (written vs. spoken)	-0.184	0.10	59.8	-1.78	0.080
Type 2 × Mod.	0.212	0.14	55.7	1.49	0.142
Type 3 × Mod.	0.134	0.14	55.9	0.94	0.350
Type 4 × Mod.	-0.062	0.14	57.1	-0.43	0.667
Type 5 × Mod.	-0.163	0.16	77.2	-1.05	0.299
Type 6 × Mod.	-0.078	0.17	93.1	-0.47	0.636
Type 7 × Mod.	0.358	0.17	101.5	2.12	0.037
Type 8 × Mod.	0.308	0.16	88.8	1.89	0.062

3.1. Gender and voice effects on acceptability

The different mean acceptability ratings for each complementizer by extraction and speaker voice is shown in Figure 3.

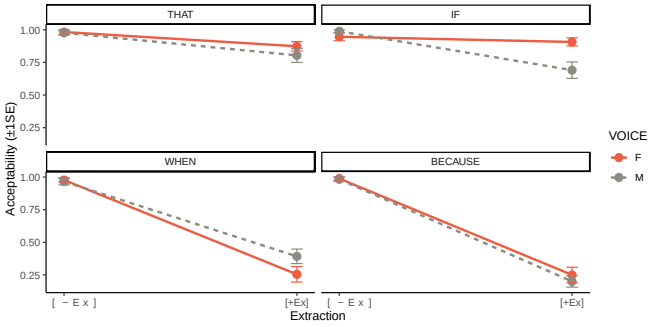


Figure 3: *mean acceptability rating (across items and participants) by complementizer [±Extraction] for female (F) and male (M) speaker voice.*

To test for potential interaction effects on the binomial acceptability between (participant) gender, (speaker) voice, and type (1–8, cf. Figure 2), we subjected the data to a mixed-effects model using the *lme4* package for *R* [21] with gender × voice × type as predictor and random intercepts for participants and items, and random slopes for gender, voice, and type for participants, and random slopes for gender and voice for items. None of the possible interactions with gender or voice nor any of their fixed 'main' effects (except type, see Figure 2) were near being significant (voice: $p > 0.8$, gender: $p > 0.9$; all interactions: $p > 0.35$) – except the interaction between voice and type 6 (*if* [+Ex]): Speakers found extraction from *if*-clauses more acceptable when spoken with the female voice than with the male voice ($p < 0.045$). This is also visible from Figure 3 (note the overlapping error bars for ±1SE). This difference removes the

© 2024 Anne Mette Nyvad et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

cost of extraction for the female voice such that extraction from *if* only has an effect for the male voice.

4. Results: RT

4.1. The overall pattern for RT

Figure 4 shows the overall mean RT for each clause type with and without extraction. The data was fitted to a mixed effects model with RT as a continuous outcome variable and type as predictor (with sliding contrasts) and random intercepts for participant and item, and random slopes for type by participant and type and trial by item. The summary of the results is provided in Table 3.

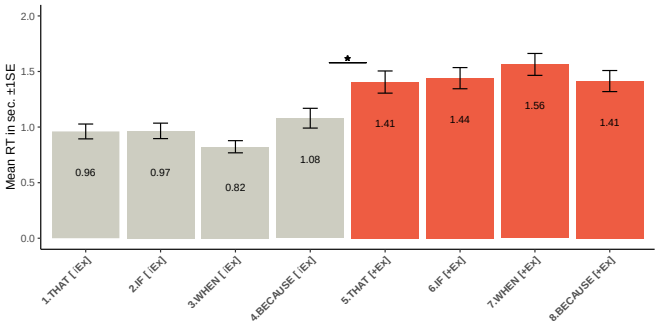


Figure 4: *RT (averaged across items and participants) by clause type. Sliding contrasts.*

Table 3: *summary of the mixed-effects model for RT with sliding contrast.*

	Estimate	SE	df	t	p
Type 2-1	-0.024	0.21	10.0	-0.12	0.910
Type 3-2	-0.072	0.18	9.9	-0.41	0.689
Type 4-3	0.202	0.16	13.8	1.24	0.237
Type 5-4	0.449	0.20	16.0	2.29	0.036
Type 6-5	0.125	0.18	14.4	0.70	0.493
Type 7-6	-0.103	0.18	16.0	-0.57	0.580
Type 8-7	-0.040	0.24	8.5	-0.16	0.874

4.2. Gender and voice effects on RT

The different mean RT for each complementizer by extraction and speaker voice is shown in Figure 5. There seems to be a small tendency for participants responding slightly faster to the female voice for *if*, whereas it is the other way around for *that* and *because*. However, the differences are very small, and the error bars overlap, suggesting non-significance.

To test whether there were any potential significant interaction effects on RT between (participant) gender, (speaker) voice, and type (1–8, cf. Figure 2), we applied a mixed-effects model using the *lmerTest Package* for *R* [20] with gender × voice × type as predictor and random intercepts for participants and items, and random slopes for gender, voice, and type for participants, and random slopes for gender and voice for items. None of the possible interactions with gender or voice nor any of the fixed 'main' effects (except type, see Figure 4) were near being significant (voice: $p > 0.91$, gender: $p > 0.38$; all interactions: $p > 0.14$). In short, none of the differences between the two voices in Figure 5 were significant.

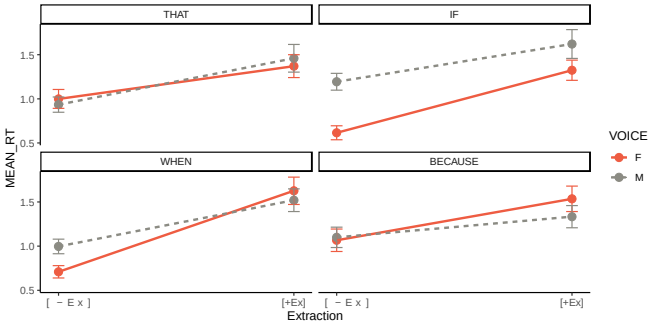


Figure 5: *mean RT (across items and participants) by complementizer [±Extraction] for female (F) and male (M) speaker voice.*

5. Discussion and conclusion

We had five research questions:

Q1: What is the overall acceptability pattern? Overall, there was no difference between extraction from *that*- and *if*-clauses, and hence, no island effect for *if*. The island effects for *when*- and *because*-clauses were both significant, but they were not different from each other (no difference between types 7 and 8). This pattern replicates the findings from Nyvad et al. [13].

Q2: Is the overall acceptability pattern for the spoken stimuli different from the pattern for the written stimuli? All but one of the possible interaction effects were not significant. Only extraction from *when* was rated slightly higher in the written modality ($p < 0.04$). This is interesting because all things being equal, such extractions are generally expected to occur more frequently (and to be more acceptable) in speech than in writing.

Q3: Any (participant) gender or (speaker) voice effects on acceptability? There were no significant interactions between participant gender, speaker voice, and type, except for type 6 (*if* [+Ex]), which was rated more acceptable with the female voice and thus only showed extraction effects in the male voice. However, our study only involved two voices (one male and one female), and it is possible that the difference is due to accidental properties of the voice or the AI-generated sequences.

Q4: What is the overall RT pattern? The only significant difference is between [+Ex] and [-Ex] ($p < 0.04$). That is, as predicted, the processing cost associated with extraction leads to an increase in RT.

Q5: Any (participant) gender or (speaker) voice effects on RT? No, there were no significant interactions or main effects of (participant) gender or (speaker) voice on RT.

One obvious weakness in this study is the number of voices involved. There was only one male and one female voice, and consequently, we must be very careful not to generalize to the population from the single significant voice effect in our study (extraction from *if*-clauses was more acceptable with the female voice than with the male voice, $p < 0.045$, removing the cost of

extraction from *if* for the female voice). Further studies with more voices are required to address this issue.

The main finding is, in fact, basically a null result, namely, that with one exception (i.e. that extraction was rated slightly higher in the written modality, $p \leq 0.037$), there was no significant effect of modality (comparing the present study to Nyvad et al. [13]). That is, the pattern of acceptability of extraction from adjunct clauses headed by *if*, *when*, and *because* is stable across the spoken and written modality.

6. Acknowledgments

The authors would like to thank the organizers of and the audience at the *Aarhus International Conference on Voice Studies*, August 17–18, 2023, Aarhus University, Denmark, the editors and an anonymous reviewer for constructive criticism. The research was funded by the Independent Research Fund Denmark (grant ID: DFF-9062-00047B).

7. References

1. Szabolcsi, A., & Lohndal, T. (2017). Strong vs. weak islands. In M. Everaert & H. van Riemsdijk (Eds.), *The Wiley Blackwell companion to syntax* (2nd ed., pp. 1–51). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118358733.wbsyncom008
2. Ross, J. R. (1967). *Constraints on variables in syntax* [PhD dissertation, MIT].
3. Huang, C.-T. J. (1982). *Logical relations in Chinese and the theory of grammar* [PhD dissertation, MIT].
4. Chomsky, N. (1973). Conditions on transformation. In S. R. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232–286). Holt, Rinehart and Winston.
5. Sabel, J. (2002). A minimalist analysis of syntactic islands. *Linguistic Review*, *19*(3), 271–315.
6. Christensen, K. R., Kizach, J., & Nyvad, A. M. (2013a). Escape from the island: Grammaticality and (reduced) acceptability of *wh*-island violations in Danish. *Journal of Psycholinguistic Research*, *42*(1), 51–70. https://doi.org/10.1007/s10936-012-9210-x
7. Christensen, K. R., Kizach, J., & Nyvad, A. M. (2013b). The processing of syntactic islands – An fMRI study. *Journal of Neurolinguistics*, *26*(2), 239–251. https://doi.org/10.1016/j.jneuroling.2012.08.002
8. Christensen, K. R., & Nyvad, A. M. (2014). On the nature of escapable relative islands. *Nordic Journal of Linguistics*, *37*(1), 29–45. https://doi.org/10.1017/S0332586514000055
9. Bondevik, I., Kush, D., & Lohndal, T. (2020). Variation in adjunct islands: The case of Norwegian. *Nordic Journal of Linguistics*, *44*(3), 1–32. https://doi.org/10.1017/S0332586520000207
10. Müller, C. (2017). Extraction from adjunct islands in Swedish. *Norsk Lingvistisk Tidsskrift*, *35*(1), 67–85. http://ojs.novus.no/index.php/NLT/article/view/1399
11. Kush, D., Lohndal, T., & Sprouse, J. (2018). Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory*, *36*(3), 743–779. https://doi.org/10.1007/s11049-017-9390-z
12. Kush, D., Lohndal, T., & Sprouse, J. (2019). On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, *95* (3), 393–420. https://doi.org/10.1353/lan.2019.0051
13. Nyvad, A. M., Müller, C., & Christensen, K. R. (2022). Too true to be good? The non-uniformity of extraction from adjunct clauses in English. *Languages*, *7*(4), 244. https://doi.org/10.3390/languages7040244
14. Lindahl, F. (2017). *Extraction from relative clauses in Swedish* [PhD dissertation, University of Gothenburg]. https://gupea.ub.gu.se/handle/2077/51985
15. Teleman, U., Hellberg, S., Andersson, E., Christensen, L., & Svenska Akademien. (1999). *Svenska Akademiens grammatik*. Svenska Akademien. https://svenska.se/grammatik/
16. Hansen, A. (1967). *Moderne Dansk. Bind I.* Grafisk Forlag.
17. Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017
18. Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y
19. R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
20. Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13
21. Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes* [Computer software]. https://cran.r-project.org/package=lme4
22. Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer New York. https://doi.org/10.1007/978-0-387-21706-2
23. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* [Computer software]. Springer. https://ggplot2.tidyverse.org
24. Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, *88*(1), 82–123. https://doi.org/10.1353/lan.2012.0004
25. Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, *34*(1), 307–344. https://doi.org/10.1007/s11049-015-9286-8

© 2024 Anne Mette Nyvad et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-010

For a systematic semantic classification of imitative prosody

Antoine Tholly[1]

[1]Sorbonne Université (STIH), France

antoine_tholly@hotmail.com

ABSTRACT

Imitative prosody (prosody miming word meanings) is an emerging field of study: associations between prosodic features and meanings is now well attested (Nygaard & al [1], Perlman & Cain [2]), and some theoretical explanations are given in terms of gestures (Fónagy [3], Perlman & al [4]), metaphors (Nygaard & al [1]), or blending (Auchlin [5]). In this paper, as an update of a previous work [6], we are working on the integration of the imitative function into the general prosodic system, the syntagmatic analysis of utterances, and theoretical linguistics (at this stage, mainly analogical and structural linguistics). In particular, we introduce semantic features, semantic chains and polysemy in the prosodic field.

Key words: prosody, sound symbolism, semantics, analogical linguistics, interpretative semantics, biological codes

1. Theoretical models: interpretative semantics and analogical linguistics

In its approach to *imitative prosody*, this paper is entirely theoretical. According to our framework, meant to distinguish different imitative and non-imitative prosodic signs, the motivation of forms leads to polysemy, which leads to contextual distinction and the consideration of three levels of paradigmatic classification (already established by the theory of interpretative semantics for verbal language). To simplify the discussion, we will refer to archetypal theoretical examples (i.e., "obvious" and "invented" examples). To listen to a few hundred empirical examples in French[1] (and no doubt soon in English[2]), classified according to the concepts established here, the reader can refer to our works [6] and [7].

Imitative prosody is made of particular *prosodic signs*, i.e. linguistic units associating meanings with suprasegmental (not phonematic) forms. Two theories help us address our subject. *Interpretive semantics* theory [8, 9] provides an account of the conditions of interpretation of the prosodic signs chosen by a speaker in an utterance, and proposes a classification of them among a network of signs. *Analogical linguistics* [11] helps to organize the typology of motivational facts (= that which is not arbitrary), which are at the core of the functioning of imitative prosody, if not of prosody in general.

1.1. The category levels and semantic chains of interpretative semantics

We refer to *interpretive semantics* to codify the interpretation of a prosodic sign in accordance with its context of appreciation, i.e., the elements co-present in the utterance (*syntagmatic* approach), and in accordance with the associated elements in the subject's linguistic competence (*paradigmatic* approach). This theory, developed mainly by Rastier (cf. [9] for an introduction in English), is in the line of the structural linguistics of Saussure and Greimas, and we have extended it to the study of prosody (and in particular the prosodic signs associated with verbal content, primarily imitative). Rastier's project is to describe in semantic features the conditions of interpretation (a semantic feature to mark the regrouping of signs in a certain domain, for example). These semantic features are used to describe signs in their categories (on the paradigmatic axis), as well as signs in their associations on the utterance chain (on the syntagmatic axis). In addition, they can be used starting from the semantic feature itself (from the concept to the various signs containing it: *onomasiology*) or starting from the signs (from the signs to the concepts they express: *semasiology*, in particular to address the problem of *polysemy*). In our view, this framework makes it possible to organize a systematic description of imitative prosody, if not of prosody in general. The metalanguage of this theory will be slightly reduced.

The *lexicon* is defined at different paradigmatic levels. The minimal level is the *taxeme*, the small class of immediate

[1] Our doctoral dissertation combined an empirical and theoretical approach, and the categories were established on the basis of 200+ examples by professional French actors declaiming poetry.

[2] The two languages operating almost or totally in same way for the facts considered here, these facts being highly motivated.

© 2024 Antoine Tholly. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

alternative oppositions (e.g., *daughter* vs. *son*). A *domain* contains a number of taxemes: these domains are paradigms of greater generality, bringing together lexical diversity under a theme. They are similar to the domain of dictionaries, e.g., *spatiality* or *politics*.[3] Finally, *dimensions* are paradigms regrouping signs according to fundamental (often dichotomous) oppositions in a given language (usually of grammatical importance): for example, for verbal language, *singular* vs. *plural*, *potential* vs. *irrealis*, *concrete* vs. *abstract*, *animate* vs. *inanimate*. We will progressively explain how these categories can be transferred to prosody (and the advantages of such a transfer). The three paradigms are exemplified in the figure below, taking into consideration a few lexical words, one of which with regards to its polysemy (cf. *knife*).

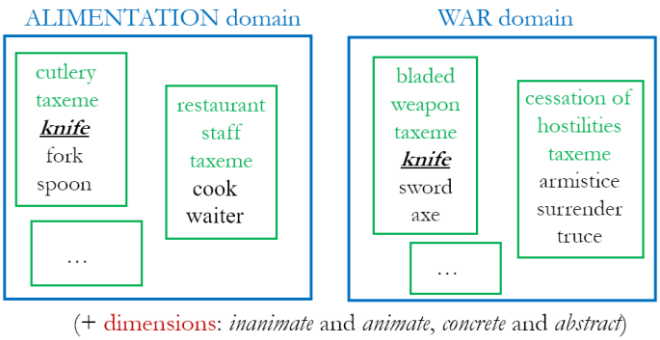


Figure 1: *taxemes, domains and dimensions paradigms*.

A final notion must be added in addition to these three paradigm levels: *isotopies*. An isotopy refers to the repetition of the same semantic feature across two signs in close proximity on the syntagmatic axis, enabling their cohesion through this semantic chain. Thus, if *knife* is co-present with *cook*, a domain isotopy of *food* is established, while if it is co-present with *surrender*, the domain isotopy established is that of *war*. Isotopy thus makes it possible to select, for a polysemous term, the semantic profile that is consistent with the context (i.e., how *knife* should be understood).

1.2. Analogical linguistics

Under the notion of *analogy*, a faculty considered primordial in the light of advances in cognitive science [10], this linguistic theory, developed mainly by Monneret [11], aims to systematically describe the links of *similarity* (or resemblance) that can exist between linguistic elements of different kinds. These different kinds of elements can be phonetic and phonological units or features, linguistic meanings (the signified and its semantic features), concepts and referents. These links of similarity can be *homogeneous* (between two phonological units, for example), or *heterogeneous* (between a phonological feature and a semantic feature, for example). This framework, merely sketched out here, offers the possibility of a more systematic understanding of *motivational* facts, i.e., non-arbitrary relations between elements of form and elements of meaning (i.e., *heterogeneous* analogies). *Homogeneous* analogies, in particular those involving categorization (shared semantic feature between two signs of the same paradigm), and cohesive relations (shared semantic feature between two signs by isotopy), will however be addressed using the theoretical framework of interpretive semantics.

2. Morphosemanticism: the motivation of prosodic forms

Within this framework of analogical linguistics, motivational facts are systematically identified and confronted. *Motivation* refers to that which is not arbitrary in the form-meaning relationship, and this broad field of study encompasses *sound symbolism*. However, as Monneret emphasizes [12, p. 8], sound symbolism is not only *iconic*, it can also be *indexical*. Indexicality and iconicity were defined by Peirce [13], then made more accessible to linguists by Jakobson [14]. According to Jakobson's definition, while iconicity refers to a relationship of suggestive resemblance between form and meaning, indexicality refers to a relationship of "existential contiguity" of one to the other: for example, the form of a spontaneous cry is motivated by the emotion felt, its meaning, and this relationship is one of physiological necessity, not resemblance. As far as prosody is concerned, indexicality has been extensively studied by Ohala [15] and Gussenhoven [16], under the designation of *biological codes*, and for vocal emotions by K.R. Scherer and colleagues [17].

2.1. Indexicality and biological codes

The fundamental argument of *biological codes* is that some facts of form-meaning motivation are identically found in the prosody of many of the world's languages and can be attributed to *physiological* constraints. Counter-examples do not contradict this so-called natural tendency, and Gussenhoven acknowledges [16, p. 80] that linguistic evolution can distend or cancel out this motivational relationship. The code relating to the *breathing* constraint[4] associates high breath level, and consequently high melody, with the notion of rupture or novelty (paragraph attacks, non-active topics), and conversely low melodic level with completion (paragraph endings [16, 18]). A continuity at the beginning of a group will thus tend to be low (not a thematic break, e.g., topics already active [19]), while a continuity at the end of a group tends to be high (not a completion [20]). The *effort code* explains the tendency of many semantically intensive prosodic signs (emphasis [20], contrastive focus [21], melism [22], vocal excitement [17]) to resort to melodic excursions and increases in loudness (leaving aside tempo, which is fast for vocal excitement, and slow for emphasis). The *frequency code*, associated with attitudes encountered in the animal world (prey/predator and seduction/rejection relationships [15]) explains a variety of affects (low-pitched negativity and authoritativeness, high-pitched positivity and fragility), and in a more *grammaticalized* way,

[3] Prosody essentially plays a grammatical or affective role. It is when it imitates a verbal content that it indeed performs a thematic, "lexical", role (expressing high spatial position, large size, etc.). These imitative domains (spatiality, size) are the domains of sound symoblism, and are generally not the same as the domains listed in dictionaries for verbal language.

[4] Speech is linked to an exhalation phase: the lungs usually project more air at the beginning than at the end of the emission, and air pressure is correlated with melody and loudness.

certain sentence modalities (high-pitched questions and low-pitched assertions [16]).

In our view, many *imitative* prosodic signs are also derived from these biological codes. More precisely, in many cases, these imitative signs function as a kind of repurposing of other *non-imitative* prosodic signs. According to the *effort code*, levels of emotional arousal/excitement are physiologically motivated in their form-meaning relationship [17]: acceleration, increased melodic level and range, and increased loudness are common to different emotions with a high level of arousal, such as *hot anger*, *distress*, etc. These features may also be reused to imitate words expressing concepts of fast movement, or brute force, or agitation. Similarly, again according to the *effort code*, prosodic emphasis, whose main role is grammatical (insistence function), may also imitate different intensity-related concepts (but not all of them: for example, large size, and not large intelligence). According to the *frequency code*, positive vs. negative qualifications, motivated in their form (breathy and high-pitched voice vs. tensed and low-pitched voice) can be reused to imitate words relating to these oppositions (e.g., *heaven* vs. *hell*)[5]. This is what we shall try to organize into categories and conditions of interpretation. However, the motivation for these imitative functions is only partially *indexical*, and also relies on *iconicity*.

2.2. Imagic iconicity

According to Peirce's definition (i), image iconicity refers to the resemblance of a sign to the object it targets (another sign) [13], whereas according to Jakobson's definition [14] this process of resemblance is established between the form and the meaning of a sign. In the latter case, we suggest that it can be (ii) a general resemblance: the sign's form perceptually evokes the sign's meaning, or (iii) a partial resemblance: within the prosodic sign, the form evokes part of the meaning. On this basis, we formulate three types of imagic iconicity for imitative prosody. We illustrate them with the *emphasis* prosodic sign, whose most expressive form combines the features of melodic excursion, volume increase and deceleration. A first kind of imagic iconiciy is (i) imitation proper: the prosodic sign evokes the *object* it targets (e.g., an emphasis positioned on *elephant* is interpreted as a vocal image of this word, i.e. a sensitive representation of this object through one of its essential characteristics, its large size). (ii) this implies, *internally to the prosodic sign*, that the prosodic form of emphasis is capable of signifying *big size*, which it does in the mode of a suggestive form-meaning resemblance. Only a few domains are imitable by this form: *biiiig* and *strooong* carry an image in their respective domains, *size* and *strength*. By contrast, *smaaaart* and *hoooot* are grammatically intensified (*very smart, very hot*), but the domains *intelligence* and *temperature* are not the object of an imitation (i.e., the prosodic form, its enlarging effect, does not give a sensible feeling of heat like it does for size). (iii) more fundamentally, and arising from the *effort code* seen above, the form of emphasis sensitively evokes the very general concept of great intensity. It is this intensive semantic feature that can be transferred to various non-imitative meanings (grammatical: insistence or intensification) and various imitative meanings (great size, imposing strength, strong sonority, etc.). The figure below summarizes these three combined distinctions (imitation of a word as the targeted object, vs. imitation of an intense quality in certain thematic domains, vs. imitation of intensity as such by the intense form).

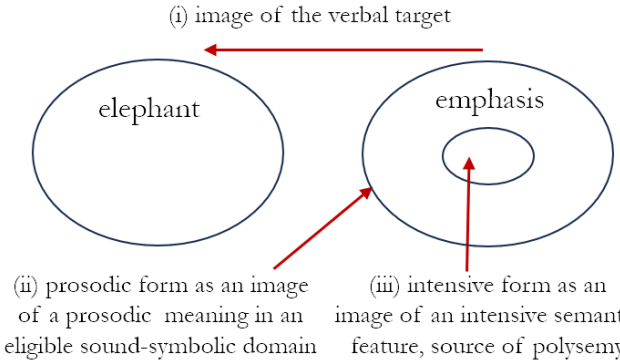


Figure 2: *combination of three types of imagic iconicity*.

2.3. Diagrammatic iconicity

Following Jakobson, while *imagic* iconicity designates a link of resemblance internal to the sign (between its form and its meaning), *diagrammatic* iconicity refers to a form-meaning resemblance between several signs [11, 14]. Given two signs, each with a form and a meaning, the relationship between their forms (of gradation, for example) is evocative of the relationship between their meanings (also of gradation). We propose four types of diagrammatic iconicity that may be of interest to the functioning of prosody.

Table 1: *four types of diagrammatic iconicity*.

Equative	=/=	Phon= ↔ Phon= ≈ Sem= ↔ Sem=
Privative	Ø/+	PhonØ ↔ Phon+ ≈ SemØ ↔ Sem+
Scalar	+/++	Phon+ ↔ Phon++ ≈ Sem+ ↔ Sem++
Polar	–/+	Phon– ↔ Phon+ ≈ Sem– ↔ Sem+

The table reads as follows: in an *equative* diagram, for example, partially equal forms evoke partially equal meanings. This is the case for prosodic dimensions (cf. 4.1.), derived from biological codes (cf. 2.1.): e.g., various signs expressing the dimension of *newness*, or the dimension of *intensity* all share part of their form. Let's take again the example of *emphasis*. Its motivation has already been examined according to the *biological code of effort* and as an *imagic iconicity*. Its motivation can also be examined according to these *four types of diagram*: (i) According to an *equative* diagram, emphasis shares some of its vocal features with other prosodic signs belonging to the intense dimension, such as contrastive focus or vocal excitation. (ii) According to a *privative* diagram, emphasis and its audible form is opposed to absence of

[5] We differentiate between a (positive or) negative *qualification* (a noun, such as *project*, is modified by an adjective: *a bad project*, or by its prosodic equivalent), and a (positive or) negative *imitation*: an inherently negative word, such as *hell*, is ornamented by a negative prosodic sign, the purpose of which is to form an expressive image of the word.

© 2024 Antoine Tholly. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

emphasis and its null form in the *French accentual group scheme* (cf. *accent initial* [20]). (iii) According to a *scalar* diagram, emphasis is opposed to more extreme emphases, which is reflected directly in the forms (*GOAAL* vs. *GOOAAAL*). (iv) According to a *polar* diagram, relevant to imitative prosody, emphasis opposes attenuation (vocal effort for *BIG*, and vocal attenuation for *small*). It should be noted that these diagrams are of the greatest interest for *categorization*: two (or more) signs of a paradigm are motivated in their relations, and this will contribute in the following sections to justifying the existence of prosodic *dimensions* and *taxemes*.

3. Semasiology: from signs to meanings

Paradigmatic semasiology lists the different meanings of the same prosodic sign, where imitation appears as one of the possible functions, and indeed where several types of imitation can emerge from a unique sign (e.g., the large size, large strength, large presence, etc. imitative meanings of emphasis). *Syntagmatic semasiology* describes the conditions of actualization of these different possibilities, which can also amalgamate (cumulate). *Onomasiology* describes the conceptual proximity of a prosodic sign to other signs, via recourse to the same dimension, domain or taxeme (*paradigmatic onomasiology*), or via isotopic links with conceptually similar verbal content (*syntagmatic onomasiology*).

3.1. Paradigmatic semasiology: one sign, several classes

One of our main concerns is to describe the *imitative function* of prosody economically and in relation to the *general functioning* of prosody, rather than postulating *ad hoc* imitative forms (lists associating a form with a meaning), without association links with the same form (*polysemy* or *homophony*), or other meanings (*categorization*). The association links are notably made with vocal signs that are not imitative, but are more fundamental to the functioning of prosody, i.e. signs with grammatical or modal/attitudinal meanings (cf. next section). These differences in interpretation need to be codified. Let's take again the example of emphasis, a single polysemous sign, and categorize its different meanings (its different profiles in polysemous relation). The same *taxeme*, or minimal paradigm of oppositions, contains *emphasis* and *attenuation* (cf. 2.3 *diagrammatic iconicity*), due to the motivation of their respective semantic cores (cf. 2.1. *biological effort code* and 2.2 *imagic iconicity iii*). The different *polysemous* imitative possibilities of these two signs are distinguished from each other by projecting into different *sound symbolic domains* eligible for the imitation of verbal content (cf. 2.2 *image iconicity ii*). These domains are not, of course, the same as the lexical domains (cf. 1.1): they are the domains in which sound symbolism is known to operate (size, strength, etc.), and within which these two imitative prosodic signs realize specific values (such as *small size* imitation for attenuation, *large size* imitation for emphasis).

The domains proposed for this pair of signs, generally attested by studies on sound symbolism (cf. references in the abstract), are the *size* domain (where emphasis can, for example, imitate the word *tall* and attenuation the word *short*), *strength* domain (e.g., for imitating the words *strong* vs. *weak*), loudness domain (e.g., for imitating the words *noisy* vs. *silent*), *quantity/weight* domain (e.g., for the words *a lot* vs. *a few*), *presence* domain (e.g., *appear* vs. *disappear*), *resistance* domain (e.g., *hard* vs. *soft*). More fundamentally, emphasis also has grammatical meanings, i.e. non imitative meanings that do not belong to any domain (but to a general intensive dimension, cf. next section).

Other imitative prosodic signs, belonging to other taxemes, can also join these same domains of sound symbolism, e.g. some prosodic sign belonging to the *strength* domain without belonging to the emphasis-attenuation taxeme (cf. 4.1 *paradigmatic onomasiology section*).

3.2. Syntagmatic semasiology: activation of a profile by context

If various prosodic signs (e.g., emphasis, positive appreciation, low excitation) are polysemous, we need to consider the *contextual conditions* under which certain meanings are realized in an utterance, rather than the other possible meanings derived from the same form. Three stages of analysis, of increased complexity, can be contemplated: (i) the mechanism of *selection* of a profile by the context, (ii) the possibility of an *amalgamation* (accumulation) of several profiles in the interpretation, and (iii) in the case of accumulation of profiles, the degree of *salience* of each in the interpretation.

(i) The contextual selection mechanism may be *prosodic* and/or *verbal*. Let's take the example of the strong excitement prosodic form. A single form (acceleration, increased melodic level and range, increased loudness) is common to different profiles. The selection mechanism is *prosodic* in the case of an *emotional* function: *despaired sadness*, *elated joy*, *panic fear*, and *hot anger* all partially use this same form [17], and it is other prosodic features that enable to distinguish these different emotions (e.g., a voice quality effect specific to anger). The selection mechanism is *verbal* in the case of an *imitative* function: for the same strong excitement polysemic form, the interpretation can be of *fast movement* (e.g., *he run runs runs*) or *brutal strength* (e.g., *he vehemently strikes his boxing partner*), if an appropriate verbal content allows one of these profiles to be interpreted.[6]

(ii) As a second step, we can consider under which utterance conditions different profiles (grammatical and imitative) can accumulate, i.e., depending on the word *targeted* by the prosodic sign. Let's take the example of an emphasis targeting the word *large*. The word *LAARGE* is first of all highlighted in its immediate context (*insistence* function, i.e. the word is considered a very important element of the utterance). Second, it is also *intensively modified* by the emphasis, a grammatical role synonymous to a modifier such as *very*, i.e. *very large*. And third,

[6] The verbal content is thus both the *source* of interpretation (it allows to comprehend what the polysemic prosodic sign of excitement is about: *fast movement* imitation, or *brutal strength* imitation), and the *target* of interpretation (the function of the prosodic sign is to imitate this verbal content, e.g. an imitation of the clause *he runs runs runs*).

the emphasis creates an *imitative image* of the word, a sort of ornamental function. If the emphasis is applied to *SMAALL*, the word is both highlighted (*insistence*) and intensified (modifying *small* into *very small*), but the word is not imitated (it would indeed be imitated by the opposite form of its taxeme, i.e. attenuation). Gradable words are prosodically intensifiable, but only some of them are imitable. Finally, a word like *table* cannot be intensified (not gradable) or imitated, it can only be highlighted (insistence function).

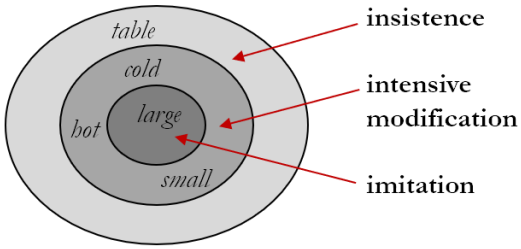


Figure 3: *amalgamation of polysemous profiles.*

(iii) In a third stage, the *amalgamation* (accumulation) approach can be refined by questioning the presumed preponderance of certain interpretations over others. In particular, we postulate a difference in treatment between *adjectives* and *nouns* (according to their prototypical semantics, i.e. not nouns derived from adjectives, such as *largeness*). Let's take the example of the voice indicating negative appreciation (tensed and low-pitched). Words like *bad* and *devil* are both imitable by this prosodic sign. The interpretation *very bad* (*intensive modification* function) may however seem more salient in the interpretation than an image of the word *bad* (*imitative* function), while the imitative interpretation of *devil* (as an imaged representation, an ornamental figure of speech) may seem more salient than the intensification function (*a terrible devil,* modifying function). This is linked to the fact that an adjective qualifies (unlike a noun), and thus this prosodic addition to the adjective amplifies the degree of qualification of the noun targeted by this adjective. For a noun, the main function seems to be to *recall* an archetypal quality of it (the *terrible* quality of *devil*, where the qualification, already known, does not appear to be added, but only shown for the sake of representation). This can be coded in terms of taxeme (minimal paradigm regrouping the pair of positive and negative signs), dimensions (for the grammatical, modification process), and domains (for the thematic, imitative process). This is examined more thoroughly in the next section.

4. Onomasiology: from meanings to signs

The advantage of a technical classification such as the one we are proposing is that it organizes closely related facts that need to be disentangled: different types of form motivation, imitative functions vs. qualifying functions, but also semantic features common to *several* different signs (*paradigmatic onomasiology*) and activated under certain contextual conditions (*syntagmatic onomasiology*).

4.1. Paradigmatic onomasiology: prosodic classification levels

It has been seen that Gussenhoven's *biological codes* [16] give rise to a variety of prosodic signs, sharing some of their form features and some of their semantic features. Similarly to the grammatical or modal/attitudinal dimensions of verbal language (cf. 1.1), prosodic *dimensions* can be proposed for these codes. These dimensions both distinguish and regroup prosodic signs. (i) These dimensions distinguish prosodic signs within their taxeme paradigm (*polar diagrammatic iconicity,* e.g., in the excitement taxeme, the distinction between the *high excitement* vs. *low excitement* pair of prosodic signs). (ii) These dimensions also regroup prosodic signs belonging to different taxemes (*equative diagrammatic iconicity*: the various high intensity prosodic signs regrouped in a high intensity dimension paradigm, and the various low intensity prosodic signs regrouped in a low intensity dimension paradigm).

Adapting [16] to the interpretative semantics theory, the *respiratory code* (also known as *production code*) leads to *completion* vs. *continuation*, *givenness* vs. *newness* dimension paradigms, the *frequency code* leads to *certainty* vs. *uncertainty*, *positive valence* vs. *negative valence* dimension paradigms, and the *effort code* leads to *high intensity* vs. *low intensity* dimension paradigms. For example, according to (i), signs belonging to the French *taxeme* of prosodic topics [19] are distinguished according to the *dimensions* newness vs. givenness, and (ii) each sign joins other prosodic or verbal signs according to these respective semantic dimensions (e.g., prosodic and verbal signs of newness)[7]. It has also been seen that *imitative* prosodic signs (e.g., word imitation through emphasis) can be cumulated with a grammatical meaning belonging to one of these dimensions (e.g., the *high intensity* dimension for the insistence and intensification functions of emphasis).

Secondly, in our work [6], we organized (and exemplified) 30 types of *imitative meanings* in 10 *domains* of sound symbolism: *movement*, *strength*, *quantity*, *size*, *sonority*, *presence*, *resistance*, *spatial position*, *temporality,* and *valence*. A significant proportion of these imitative signs are also projections of *taxemes* (e.g., the emphasis vs. attenuation taxeme, or the strong vs. weak excitement taxeme) in different *domains* (e.g., *movement* and *strength* for the pair of signs of the latter taxeme, as a result of polysemy). These domains host either prosodic signs from different taxemes or isolated prosodic signs (not belonging to a minimal paradigm of definition). For example, in the *strength domain*, the imitation of *brutal strength* (loud and fast, projected from the excitement taxeme to this domain), is opposed to the imitation of *stable strength* (loud and slow, projected from the emphasis-attenuation

[7] Consider this theoretical example: Xa and Xb from taxeme X, and Ya and Yb from taxeme Y. As a reminder, taxemes group together alternative choices, such as different kinds of prosodic focus, different kind of prosodic paragraphs, etc. Our point is that Xa and Ya (from different taxemes) can also have a common origin, through a biological code, which continues to manifest itself synchronically through a common grammatical feature (of continuation, of intensity, etc.): this is what we call a *dimension*, grouping together signs according to a high-generality paradigm (and in these cases of motivation, the grouping is also reflected in their formal features: cf. diagrammatic iconicities). The third paradigm, i.e. *domain*, concerns the almost "lexical" uses of prosody, i.e. when prosody assumes a thematic-imitative function (space, size, etc.).

© 2024 Antoine Tholly. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

taxeme), and to the imitation of *solemn strength* (loud and low-pitched, derived from Ohala's *frequency code* of animal communication).

4.2. Syntagmatic onomasiology: isotopy relations

Imitation can occur if the same meaning can be assigned to both the *imitated* verbal content and the *imitating* prosodic content. First of all, the two signs (verbal and prosodic) must be concomitant on the syntagmatic axis: the prosodic sign physically targets the verbal content, as suprasegmental features "over" the segmental pronunciation of the verbal content (i.e., another kind of *indexicality*). Secondly, not all of the meaning is identical: when an emphasis imitates the word *elephant* by positioning itself on the word (*I saw an ELEPHAANT*), it is only certain semantic aspects of the word that are selected, i.e., its large size, and not the fact that it is an animal, with a trunk, etc. To account for this, certain identical meanings of these different signs (*onomasiology*) are captured on both sides by an *isotopy* relationship (*syntagmatic* connection of two signs via the partial identity of their meanings [8]). Let's get into more technical detail. First of all, interpretative semantics distinguish between *semantic features* that relate to membership in a paradigm (*taxemes*, *domains*, *dimensions*), and those that relate to a distinction between signs within the paradigm (semantic *distinctive features*). Secondly, the theory distinguishes between their *inherent* status (actualized in all contexts) and *contextual* status (needing to be triggered through an isotopic connection with another sign). For our subject, inherence is justified by the *motivational* facts set out in 2.: emphasis, for example, is recognized by its grounding in several resemblance iconicities, and in a biological indexicality. Its *distinctive* /+intense/ *feature* is thus recognized in all contexts. On the other hand, the *domain* to which this intensity belongs (*size*, *sonority*, *strength*, etc.) depends on the context (and in most contexts, this emphatic core plays no imitative role, but only a grammatical one of insistence or intensification). To formalize this imitative interpretation, we therefore postulate a double isotopic relationship, at different levels in the sign's semantic structure, and with a different relationship to context.

The figure below illustrates one of these imitative relations to be coded as a dual identity relationship between the imitated word and the imitating prosodic sign, namely a /+intense/ isotopy and a size-domain isotopy. On the prosodic side, emphasis and attenuation are *distinguished* by intensity features (/+intense/ vs. /-intense/), and they are regrouped by a *domain* feature (here: the domain of size). These two prosodic signs contextually express *large size* vs. *small size*. This contextual meaning of emphasis (i.e. *big size*) is thus broken down into two semantic features: a *distinctive* /+intense/ feature, interpreted *inherently*, as the result of iconic and indexical motivational facts, and another feature related to a sound symbolism category, that of size. This second semantic feature (i.e. size domain) depends on context: indeed, emphasis can be imitative in other domains, such as strength or sonority, in addition to its grammatical roles (which do not belong to thematic domains but to dimensions). As a result, the domain feature can only be interpreted if an *isotopy* takes place contextually (here: between *emphasis* and the word *elephant*, where the latter can also be defined contextually as /+intense/in the size domain, i.e. as *large*).

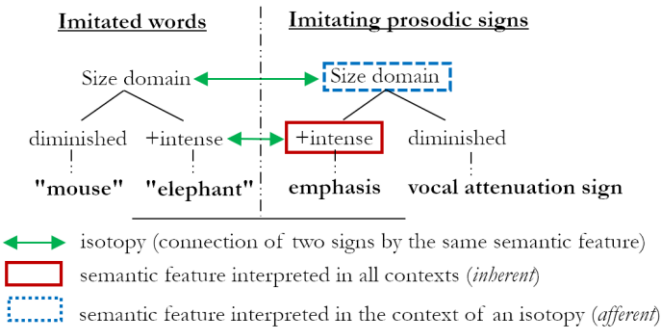


Figure 4: *imitation relation, coded by a double isotopy*.

This semantic analysis can be extended to the imitated verbal content: for a given word, to what extent are its *distinctive* semantic feature (such as +intense) and its *domain* feature (such as size) recognized independently of their isotopic relations with the prosodic content that imitates this word? In other words, what is the interpretive distance between the definition of *elephant* in an ordinary context (e.g. in a dictionary), and *elephant* as captured in the imitative relation, i.e. as an archetype of large size (coded as /+intense/ in the size domain)?

5. Conclusion: limits of the proposed classification

Our modeling of imitative prosody, according to the stated conditions (categorization that potentially crosses the entire prosodic field, disentanglement of motivational types, discourse activation by semantic chains) remains a work in progress. One of the first points of improvement envisioned is to open up to other semantic theories, and notably (presumably) cognitive semantics.

Another point of concern is the combination of the theoretical and the empirical. Readers will surely have recognized our Saussurian heritage, which, by methodology, largely withdraws linguistic thinking from the realms of substance (phonetics). As *imitative signs* function in an impressive way (*image* production), different phonetic forms can achieve the same goal, with flexibility in the degrees of realization (depending on the *degree* of salience the speaker wishes to give to the image). For example, for emphasis, the enlargement effect of the sound print can work through a mixture of melodic excursion, loudness increase, and deceleration (syllabic lengthening). Not all vocal features are necessarily present, creating *allomorphs*. The difficulty is how to combine, for an assumed prosodic sign, both a phonic variation (allomorphism) and a semic variation (polysemic profiles). We shall therefore conclude with two types of remark, on allomorphism on the one hand, and on homophony on the other.

(i) The various possible vocalizations of an imitative sign (its allomorphs) must be studied empirically. A given imitative meaning (e.g., the imitation of a word expressing intense sonority) can be considered as one of the possible profiles of a polysemous imitative sign (in this case, emphasis), if it uses the same allomorphs as the sign's other profiles. If, on the other

hand, the imitation of an intense sonority (e.g., an imitation of the word *gong*) only uses the loudness increase parameter (*a GONG*), and not the other allomorphs of emphasis (?*a gooong*), it would seem preferable to consider it as a *different sign* than the emphasis. For instance, the imitation of length (*looong*) in different domains (*time*, *space*, *size*) only uses a syllabic lengthening and not an increase in loudness (**LONG*): this meaning should *not* therefore be considered as one of the profiles of emphasis (this sign does, nevertheless, amalgamate with grammatical insistence and intensification functions, cf. 3.2. However, this is an accumulation of different signs, not an accumulation of profiles of the same polysemous sign).

(ii) It has been shown that imitative profiles can be polysemous with non-imitative profiles (emphasis for *large size* vs. emphasis for *intensification*, for example). In addition, these imitative signs can also be *homophones* of other prosodic signs. In the case of the final lengthening allomorph of emphasis, for example, other signs also use this form: as a phonological feature of a major prosodic boundary [20], as a positive-lyrical prosodic sign (*a poeeet*), or as a hesitation mark. However, as we have seen in 3.2. (cf. *syntagmatic semasiology* section), different meanings can *accumulate* in the interpretation of the same form. The use of the notion of amalgam is in line with our aim of not separating the study of prosodic imitation from the rest of the prosodic field, and it is up to empirical studies of a *psycholinguistic* type to determine, for each context, to what extent certain functions are more *salient* than others in interpretation, when they are accumulated in the same form.

6. References

1. Nygaard, Lynne C., Herold, Debora S. & Namy, Laura L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive science*, 33-1, 127-146.
2. Perlman, M. & Cain, A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, 14-3, 320–350.
3. Fónagy, I. (1983). *La vive voix : essais de psycho-phonétique*. Payot.
4. Perlman, M., Clark, N. & Johansson Falck, M. (2015). Iconic prosody in story reading. *Cognitive Science*, 39-6, 1348-1368.
5. Auchlin, A. (2013). Prosodic iconicity and experiential blending. In Hancil S., Hirst D. (Ed). *Prosody and Iconicity*. Amsterdam, John Benjamins, 1-31.
6. Tholly, A. (2022). Analyse sémantique de la prosodie imitative. Doctoral dissertation, Sorbonne Université. Published by HAL Science [CNRS] (tel-03948396 (https://hal.science/tel-03948396).
7. Tholly, A. (2023) [website]. Project 1 for the description and documentation of imitative prosody. (https://tholly.net/project1).
8. Rastier, F. (1987). Sémantique interprétative, PUF.
9. Rastier, F. (2015). Interpretative semantics. N.Riemer (Ed) *The Routledge Handbook of Semantics*. Routledge, 507-522.
10. Hofstadter, D. (2001). Analogy as the core of cognition, Gentner, D., Holyoak, K. J., Kokinov, B. N. (Ed) *The Analogical Mind: Perpectives from Cognitive Science*, M.I.T. Press, p. 116-144.
11. Monneret, P. (2014). L'iconicité comme problème analogique. *Le Français Moderne - Revue de linguistique française*, vol. 1, 46-77.
12. Monneret, P. (2019). Le symbolisme phonétique et la fonction iconique de l'analogie. *Signifiances (Signifying)*, *3*(1), 1-19.
13. Deeley, J. (1994). *The Collected Papers of Charles Sanders Peirce*. Electronic edition. (https://colorysemiotica.files.wordpress.com/2014/08/peirce-collectedpapers.pdf)
14. Jakobson, R. (1965). Quest for the essence of language. *Diogenes*, 13, 21-37.
15. Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, *41*(1), 1-16.
16. Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge University Press.
17. Bänziger, T & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech communication*, 46(3-4), 252-267.
18. Grobet, A., & Simon, A.-C. (2001). Différents critères de définition des unités prosodiques maximales. *Cahiers de linguistique française*, 23, 143-163.
19. Grobet, A., & Simon, A.-C. (2009). Constructions à détachement à gauche: les fonctions de la prosodie. Apothéloz, D., Combettes, B. & Neveu, F. (Ed), *Les linguistiques du détachement.* Peter Lang, 289-303.
20. Delais-Roussarie, E, Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., Yoo, H.-Y. (2015). Intonational Phonology of French: Developing a ToBI System for French. In: S. Frota, S., P. Prieto, P. (Ed), *Intonation in Romance*. Oxford, Oxford University Press, 63–100.
21. Astésano, C., Magne, C., Morel, M., Coquillon, A., Espesser, R., Besson, M. R., & Lacheret-Dujour, A. (2004). Marquage acoustique du focus contrastif non codé syntaxiquement en français. *Proc 25èmes Journées d'Etudes sur la Parole*. Association Francophone de la Communication Parlée (AFCP).
22. Caelen-Haumont, G. (2002). Prosodie et dialogue spontané: valeurs et fonctions perlocutoires du mélisme. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, 21, 13-24.

© 2024 Antoine Tholly. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-011

“A human way to talk”: poetry and synthetic voice

Vadim Keylin[1]

[1]Universität Hamburg, Germany

vadim.keylin@uni-hamburg.de

ABSTRACT

The association between the voice and identity, authenticity and selfhood, is a thread through much of Western culture. In poetry performance, this association becomes particularly prominent as the speaker inhabits the lyric subject, lending it their corporeality through their voice and imbuing the spoken with their lived experience. It is the reason behind audiences’ strong preference for an author’s own reading of their poetry compared to actors’ readings, which are often perceived as “fake”. What happens, then, when a poem is performed not by its author’s voice, nor by an actor’s voice, but by a text-to-speech (TTS) engine? This paper focuses on the case of Spoken Word poet Dave Young, aka The Shouting Mute, who does not possess biological speech and is reciting his poems in the TTS voice of his communication aid device. Subverting the expectations of the poetry reading format, Young’s performances challenge the association between authenticity, liveness, and biological voice, foregrounding the gap between the disembodied technological speaker and the silent embodied author. In doing so, they further question the status of the biological voice as a unique marker of humanity, a carrier of identity and embodiment.

Key words: contemporary poetry, Spoken Word, speech synthesis, TTS, disability.

1. Introduction

Even though poets’ voices are rarely trained and do not possess the same expressive range as actors’ or singers’ voices do, the voice is crucial to the experience of a poetry reading. As it connects the utterances of the lyric “I” to the embodied speaker and their lived experiences, the voice both imbues them with a veneer of authenticity and grounds them in the here and now of the performance situation, reaching out to the audience in what Erika Fischer-Lichte calls the “autopoetic feedback loop” [1, p. 50]. What happens, however, when the poet speaks not in their biological voice, but in a synthetic one, through a communication aid device? What implications does it have for the audience’s experience of a poetry reading? In the following, I will discuss these questionsusing the example of the British performance poet, artist, and theatre maker Dave Young, aka The Shouting Mute. Young has cerebral palsy – a condition that limits his control over his bodily movements, including those necessary to produce physical speech. Thus, to communicate and perform his poetry he uses an augmentative and alternative communication (AAC) device equipped with a synthetic voice. Examining the ways Young’s poetry performances enact authenticity and liveness despite the lack of biological voice, I will show how they challenge the entrenched assumptions about voice, identity, and poetic expression.

2. Authenticity

According to Charles Bernstein, in a poetry reading,

> explicit value is placed almost exclusively on the acoustic production of a single unaccompanied speaking voice, with all other theatrical elements being placed, in most cases, out of frame. The solo voice so starkly framed [...] physically connects the speaker and listener, moving to overcome the self-consciousness of the performance context [2, pp. 10–11].

The poet’s voice furthermore produces so-called authenticity effect: the poetic utterances are perceived as being authentic and truthful even though the audience has no recourse to assess their truthfulness [3]. On the one hand, the unspectacular and minimalistic staging of the poet’s voice invokes intimacy and intensity [4, pp. 95–96]. On the other hand, mediating between the material and the symbolic, the speaking voice makes the audience identify the subject of the utterances with the embodied speaker, projecting their identity and lived experience onto the content of the poem [3, p. 160]. Finally, authenticity is evoked by the biological voice itself due to the entrenched belief “about the voice as a cue to interiority, essence, and unmediated identity” of the speaker [5, p. 2]. As Amanda Weidman points out, Western philosophical thought generally casts the biological voice “as guarantor of truth and self-presence, from which springs the

© 2024 Vadim Keylin. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

familiar idea that the voice expresses self and identity [...] The model of the speaking subject [...] embodies all of our notions of voice as presence, authenticity, agency, rationality, will, and self" [6, p. 233]. Such discursive background frames the lyric utterances in the poet-performer's own voice as necessarily truthful, as both the lyric "I" and the voice are presumed to be indexes of their interiority.

The authenticity effect is particularly salient to Spoken Word poetry – a scene to which Young belongs – because of its general "speak your truth" ethos. As Katie Ailes notes, the expectation of poets performing autobiographical material in an authentic way is particularly strong in the Spoken Word community to the point of becoming a demand and the main criterium of aesthetic judgement [7]. This particularly affects poets from minority groups – including poets with disabilities. Some of the Spoken Word poets that Ailes had interviewed "said that they feel the pressure to write about their marginality… in order to present themselves as more 'authentic'", feeling "the need to perform their identities in a certain way to match the audience's pre-conceived notions of what they should be" [7, p. 150]. Ailes further remarks that "the performer's emotional engagement seems necessary for certain pieces to ring 'true' […] it is important that the emotional affect of the performance at least somewhat reflects the emotional tone of the poem for authenticity to be projected" [7, p. 146].

At first glance, the synthetic voice of Young's communication aid lacks the expressive apparatus of a biological voice that is necessary to produce such authenticity effect. It sounds monotonous, austere, and extremely poor in affect, hardly laying any claim to intimacy and intensity. It is a commonly held view that such voices are incapable of emotional engagement with either the audience or the poem's content, as text-to-speech technology can only sound the written text, not the underlying emotion. Brandon LaBelle argues, for example, that the synthetic voice is a mediatisation of the second order: if the acousmatic voice of or recording technology is separated from its originating speaker, then the synthetic voice has never originated in an embodied speaker in the first place [8]. According to LaBelle, the already "disembodied, radiophonic, and electronically manipulated voice is emptied of psychology, spirit, or granularity with the synthetic voice" [8, p. 161]. The synthetic voice in a poetry reading can thus be regarded as analogous to an actor's performance – in the sense of being equally removed from Bernstein's austere, yet aesthetically and emotionally charged, ideal of poetry reading. Whereas the actor's reading is typically assessed as inauthentic because of excessive dramatization [2], the text-to-speech engines go too far in the opposite direction: fit, perhaps, to express factual truth, but not the emotional one.[1]

However, I would argue that the synthetic voice of Young's poetry performances still enacts authenticity – precisely *because* it is synthetic. It can be described as a "minus-device" in Yuri Lotman's terms – an aesthetic device that draws its meaning from the absence of another aesthetic device [13, pp. 22–30], in this case, the expressiveness of biological speech. Speaking in a monotonous synthetic tone is part of Young's lived experience of a person with disability and thus attests to the authenticity of his poems more so than a "naturally" sounding voice would.

In a scene from Young's theatre show "Grow Up and Just Love Chocolate" that debuted in 2019 at Salisbury Playhouse, the poet performs his poem "The Night's Planning" together with three other actors.[2] The lyric "I" of the poem is a man with the same disability as the author preparing to go on a first date. A pivotal part of the process is coming up with "a good chat-up line to program into my communication aid", which is delivered at the very end of the scene: "Have you ever been on a silent date? / Because you're so amazing I'm speechless". While most of the poem is spoken by the actors, the chat-up line itself is read by the communication aid device. The use of synthetic voice in the scene amplifies the pun – the ambiguous meaning of "speechless" as both "unable to speak because of being overcome with emotion" and "lacking biological speech". Moreover, it gives the line emotional depth beyond its self-deprecating humour, performing the line exactly as it would sound in a real-life situation. Because the device's synthetic voice in its inexpressiveness is unable to play-act – to imitate and fake emotion – it serves as the guarantor of the emotional truth of the poem.

As Young details on his website, similar considerations drove the creation of "Grow Up and Just Love Chocolate":

> During the research and development period, The Shouting Mute initially explored actors taking over his voice and performing his poetry using different accents and vocal ranges to bring a new voice to his words. But by taking out his electronic voice completely, the performance lost an aspect of his unique voice and it became difficult to see his world. So he reintroduced his voice and performance alongside that of performers to create a dialogue within his poetry.
>
> This mix of electronic voice, physical voice and visual voices created colour, energy, and vibrancy, enabling a deeper meaning and emotion to be conveyed within the work. […] 'Grow up and Just Love Chocolate' uses a variety of techniques to express and explore the human voice. As a poet The Shouting Mute uses an alternative communication device, which allows him to question and

[1] Emotionally expressive text-to-speech is a fast developing field within current speech synthesis powered by neural network-based technologies [9], [10]. Such technologies can furthermore infer emotional register directly from the text [11] and even be designed specifically for poetry recitation [12]. Poetry performances with such systems would be largely equivalent to actors' readings (in case of [12], the system is directly trained on such recordings). While such recordings may sound barely distinguishable from biological human speech, the authenticity effect relies not so much on the perceived expressive qualities of the vocal performance, but rather on the situation of the embodied speaking and the belief that such a situation gives immediate access to the poet-performers interiority. As such, it is doubtful that emotionally expressive TTS would be significantly more capable of producing such effect than the older system (unless, perhaps, the listener has been misled that they are listening to the poet's own voice), At the moment of writing, I am not aware of poets using such technologies in their performances, making it impossible to test this empirically.

[2] See video documentation here: https://www.youtube.com/watch?v=bkIFlXWxriw

> challenge his audience about who's voice they imagine when they hear the emotionless robotic voice of his communication device. What emotions and feelings can be conveyed in monotone? [14].

The relationships of AAC users to the identities performed by their synthetic voices is complex and nuanced. On the one hand, the range of voices available to them is limited by the technology they use. Some communication producers aim to create individualised voices for their clients whose vocal quality would match the speaker's visual presentation – their age, gender, ethnicity, etc. However, these voices are expensive and most AAC users have to choose from the standard options that come with their devices.[3] Reviewing the options provided by one such communication aid, Alper [15, p. 63] observes how they disproportionally favour American white male speakers, leaving female and minority users with much poorer choice.

In his poem "Voice", Young compares the process of choosing an AAC voice to dating:

> Choosing a voice is a bit like dating.
> She is nice,
> She is sexy in that little black dress,
> She is just a bit of a cow.
> It is looking in the best sweet shop ever,
> Yum!!!
> Sometimes I am a puppy,
> Peeing everywhere with excitement.
> Choosing the best futuristic space age one that I can find.
> I pay for my voice,
> And they aren't cheap
> At least £10000 or more.
> Just for a voice but I don't mind.
> At least it's the coolest
> Voice to me,
> And you haven't got one. [16, p. 9]

The lyric subject of the poem does not identify with the chosen voice: it is externalised as distinct entity to be interacted with. However, the dating metaphor emphasises the special character of their relationship to the chosen voice: much like one's significant other, it becomes a part of their identity, while remaining separate. The last line also evokes the metaphoric meaning of having a voice, as "for AAC users, their communication aids, eye gaze, picture boards and sign languages become their voice in a far more meaningful way than their own biological speech ever could" [17, p. 3]. At the same time, AAC users also come to identify with the sound of their synthetic voices, at least to some extent. Stephen Hawking, for example, famously chose to keep his old formant-based robotic voice, despite the availability of more "natural" sounding options because it has become a part of his image and identity [18, p. 18].

Young's relationship to his voice can also be illustrated by the beginning of the video documentation of his live performance at the Tolpuddle Martyrs' Festival in 2022.[4] The first few seconds of the video show him playing with a different voice option before switching "back to [his] American Robot Twang" – the voice that he normally uses in his performances and everyday life. This ironic designation points to an identity mismatch between the voice (American) and the user (British). However, Young consciously chooses this voice over a British-sounding one that he employs ironically at the beginning of the performance. He describes the latter as "a mix of Pierce Morgan and Prince Charles" and ironically says that he just needs his pipe and his newspaper – in other words, he finds it a caricature, too stereotypically British. Despite its American accent, Young finds in the sound of his usual AAC voice more qualities that he can identify with. However, "American Robot Twang" can also be read as a term of endearment, given how often Young refers to his voice in this way – for example, in the subtitles to "The Night's Planning" or his presentation video on his website. This highlights the ambiguity of the relationship between the poet and his synthetic voice – partly identifying with it and partly treating it as a long-term companion.

Thus, much like in poetry readings by able-bodied artists, Young maintains the authenticity effect in his performances through the continuity between the poet-performer's utterances and their everyday performance of self that is guaranteed by the voice. However, rather than presenting the voice as the unique and unmediated expression of the speaker's body that embodies the truth of the spoken, it highlights the technologically conditioned performativity of everyday selves and the complexity of relationship one may have with one's self.

3. Liveness

Attached to the idea of authenticity is that of liveness – of the co-presence of the speaker and the listener in the here-and-now of the performance. Erika Fischer-Lichte speaks of an "autopoetic feedback loop" – the co-constitution of performance in live embodied interaction between the performers and the audience [1, p. 50]. This loop guarantees the uniqueness of the performance experience, its unpredictability and irreducibility to either interpretation or documentation, even though the performance itself may be to a large extent mediatised [1, pp. 67–70]. She further names voice an important means to achieving co-presence and autopoetic feedback loop:

> Through [voice], the bodily being-in-the-world of the articulating subject expresses him/herself and addresses those who hear him/her in their own bodily being-in-the-world. The voice builds a bridge and establishes a relationship between two subjects. It fills the space between them. By making their voices audible, people reach out to touch those who hear it [1, p. 130].

While for physical theatre and performance art, voice is but one of the mechanisms of embodiment, for poetry performance it is largely the only one given the genre's aversion to theatrical

[3] The recent developments in neural network-based TTS may significantly cheapen the production of such individualised voices. To what extent it will have an effect on consumer prices remains to be seen.

[4] https://www.youtube.com/watch?v=miOR_7uShUE

© 2024 Vadim Keylin. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

elements. At the same time, according to Anna Bers, mediatisation – even in such real-time formats as livestreams – robs the voice of its embodied character and transgressive potential, which in turn makes liveness and co-presence impossible [19].

Young's poetry performances, like the one at Tolpuddle Martyrs' Festival discussed above, trouble this idea of liveness. On the one hand, the communication aid needs to be pre-programmed – the poem's text has to be entered into the device in advance. It is then sounded at the performance by the text-to-speech software in exactly the same way as it would in any other time and space. In that sense, his reading of the poem is not live – even if the software operates in real time – as it is not reactive in any way to the here and now of the event.

But on the other hand, the poet-performer is corporeally present in the same space as his audience. His performances are not mediatised in the sense that Bers and Fischer-Lichte use – they are neither recordings, nor broadcasts. The "intense experience of presentness" [1, p. 96] is created already by the presence on the stage of a non-normative body that the ableist society wants hidden away as not to disturb its sanitised image of itself. But it is further reinforced by Young's expressive body language that commands the audience's attention and centres him as the subject of his poetic speech.

There is a poignancy, however, beyond the simple corporeal presence of a disabled person in the public space, to Young's performances being specifically *performances of poetry* – one that is also exemplified in the above poem. Disability scholar Bill Hughes argues that

> in the non-disabled imaginary disability is an 'ontological deficit' – a reduction of *'leib'* to *'korpor'* [*sic*], human to animal, subjectivity to flesh, identity to excessive corporeal presence. It is this deficit of credibility that provides the spurious rationale for the disposal of disabled bodies by means of elimination (*inter alia* extermination or segregation) or correction (*inter alia* sterilisation or rehabilitation) [20, p. 17].

The same discourse of western modernity that produces this dichotomy also casts speech – and particularly poetry – as a primary means of expressing, constructing, and performing subjectivity. Speech is what distinguishes the subject from the body, the human from the animal. Through the simultaneity of his poetic speech and his bodily presence as a disabled person, Young challenges the society's ableism and claims his right to selfhood and humanity.

Moreover, in poetry performances by able-bodied artists, the simultaneity of speech and corporeality is mediated by the biological voice that maintains the unity of the two. The biological voice – and, by extension, biological speech – carries in it the marks of the body that produces it, marking it in turn as the source of the utterances. Conversely, Young's performances are characterised by a stark difference – a stark contrast even – between the non-normative body and its body language and the digital speaking voice that can be described as hyper-normative. Its vocal quality is designed to be idealised: its timbre is soft and low, its diction and articulation, clear, its intonations, calm and confident. As most current synthetic voices, it does not possess any excess – neither in the form of of extreme emotionality (or any kind of emotionality), nor imperfections or paralinguistic sounds. The design of TTS is driven first and foremost by the values of naturalness and intelligibility, reflecting the ideal of the voice as a neutral carrier of linguistic meaning. Even the recent developments in emotionally expressive speech synthesis [9]–[11] do not challenge this hyper-normativity as they only implement quantifiable and normativised sets of emotions to improve the semantic precision of the TTS, avoiding or ignoring the more extreme modes of the voice.

However, the dialectic relationship between the excess of Young's expressive body language and the lack of excess in his synthetic voice highlights the non-normativity of the hyper-normative. His poetry performances challenge the normativity's claim to "natural order", revealing the inherent ableism of the synthetic speech designers' quest for "naturalness". As Alper points out, "[w]hile computer scientists and interaction designers ask questions about how human a talking computer should sound, we need to reckon with the fact that speaking through a computer is also a human way to talk" [15, p. 36].

4. Conclusion

Medical historian Katherine Ott critiques the cultural theory's penchant for using disability and prosthesis as a metaphor, whereby "any machine or technology that intervenes on human subjectivity, such as a telephone, a computer, or a sexual device, can be said to be prosthetic" [21, p. 1]. While "intellectually provocative and culturally insightful" such studies disregard the actual lived experience of assistive technology users, who "are not cyborgs or bionic beings. Nor are they merely metaphors for empire, nationhood, or modernist anomie" [21, pp. 1–3]. The material history of disability and prosthesis inverts the notion of the cyborg, revealing the human-machine symbiosis to be *just* human rather than trans- or post-human.

Similarly, Dave Young's earnest poetry performed in his "American robot twang" challenges the notions of authenticity and liveness unmasking their connection to the normative, non-disabled ideas of corporeality and subjectivity. It highlights the inherent ableism of the claim that the biological voice is the essence of selfhood and guarantor of presence. Young's reading of his poetry is no less authentic and no less live than that of able-bodied poets – perhaps, even more so in some ways, given his AAC's inability to play-act and pretend. His performances contrast the phantasmic self projected by the hyper-normative vocality of the synthetic speech with the poet's non-normative embodied self that is the source of the utterances. Drawing the listener's attention to the gap between the two, they scrutinise the very idea of humanity in the digital age.

5. Acknowledgments

This article is part of a project that has received funding from the European Research Council (ERC) under the European Union's

Horizon 2020 research and innovation programme (grant agreement No 884177).

Views and opinions expressed here are those of the author only and do not necessarily reflect those of the European Union (EU or the European Research Council (ERC)). Neither the EU nor the ERC can be held responsible for them.

6. References

1. Fischer-Lichte, E. (2008). *The Transformative Power of Performance. A New Aesthetics* (S. Jain, Trans.). Routledge.
2. Bernstein, C. (1998). Introduction. In C. Bernstein (Ed.), *Close Listening: Poetry and the Performed Word* (pp. 3–23). Oxford University Press.
3. Novak, J. (2017). Live-Lyrik. Körperbedeutung und Performativität in Lyrik–Performances. In A. Bers & P. Trilcke (Eds.), *Phänomene des Performativen in der Lyrik. Systematische Entwürfe und historische Fallbeispiele* (pp. 147–162). Wallstein.
4. Neumark, N. (2010). Doing things with voices: Performativity and voice. In N. Neumark, R. Gibson, & T. van Leeuwen (Eds.), *Voice: Vocal Aesthetics in Digital Arts and Media*. MIT Press.
5. Eidsheim, N. (2019). *The race of sound: Listening, timbre, and vocality in African-American music*. Duke University Press.
6. Weidman, A. (2015). Voice. In D. Novak & M. Sakakeeny (Eds.), *Keywords in sound* (pp. 232–246). Duke University Press.
7. Ailes, K. (2021). 'Speak your truth': Authenticity in UK spoken poetry. In L. English & J. McGowan (Eds.), *Spoken Word in the UK* (pp. 142–153). Routledge.
8. LaBelle, B. (2010). *Raw Orality. Sound Poetry and Live Bodies". Voice. Vocal Aesthetics in Digital Arts and Media* (R. Gibson, N. Neumark, & T. Leeuwen, Eds.). MIT Press.
9. Kang, M., Han, W., Hwang, S. J., & Yang, E. (2023). *ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models*. arXiv.org. https://arxiv.org/abs/2305.13831v1
10. Cho, S., & Lee, S.-Y. (2021). *Multi-speaker Emotional Text-to-speech Synthesizer*. arXiv.org. https://arxiv.org/abs/2112.03557v1
11. Mukherjee, A., Bansal, S., Satpal, S., & Mehta, R. (2022). Text aware Emotional Text-to-speech with BERT. *Interspeech 2022*, 4601–4605. https://doi.org/10.21437/Interspeech.2022-11293
12. Koch, J., Lux, F., Schauffler, N., Bernhart, T., Dieterle, F., Kuhn, J., Richter, S., Viehhauser, G., & Thang Vu, N. (2022). PoeticTTS - Controllable Poetry Reading for Literary Studies. *Interspeech 2022*, 1223–1227. https://doi.org/10.21437/Interspeech.2022-10841
13. Lotman, Y. (1976). *Analysis of the poetic text* (D. B. Johnson, Trans.). Ardis.
14. Young, D. (2018, July 6). Theatre. *The Shouting Mute*. https://theshoutingmute.wordpress.com/grow-up-and-just-love-chocolate/
15. Alper, M. (2017). *Giving voice: Mobile communication, disability, and inequality*. MIT Press.
16. Young, D. (2014). *The Shouting Mute. My Life*. Fixers. https://issuu.com/fixersdesign/docs/cp_booklet_final_v3_no_bleed
17. Radosavljević, D., Pitrolo, F., & Salazar Cardona, J. F. (2021). *The Unspoken Project—The Making of The Voice Monologues.* [Audio podcast transcript]. Auralia.Space, Royal Central School of Speech and Drama. https://www.auralia.space/wp-content/uploads/2021/02/LMYE-LABORATORY-3-The-Unspoken-Project.pdf
18. Mialet, H. (2012). *Hawking incorporated: Stephen Hawking and the anthropology of the knowing subject*. The University of Chicago Press.
19. Bers, A. (2021). '[O]ft nur ein schlechterer Ersatz'? Chancen und Grenzen der Stimme im Online-Lesungsvideo. *Wiener Digitale Revue*, *3*. https://doi.org/10.25365/WDR-03-02-04
20. Hughes, B. (2012). Civilising Modernity and the Ontological Invalidation of Disabled People. In D. Goodley, B. Hughes, & L. J. Davis (Eds.), *Disability and social theory: New Developments and Directions* (pp. 17–32). Palgrave Macmillan.
21. Ott, K. (2002). The Sum of Its Parts: An Introduction to Modern Histories of Prosthetics. In K. Ott, D. Serlin, & S. Mihm (Eds.), *Artificial parts, practical lives. Modern histories of prosthetics* (pp. 1–42). New York University Press.

© 2024 Vadim Keylin. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

DOI: 10.2478/9788366675513-012

The conversational (human?) voice of artificial intelligence

Jeanette Landgrebe[1]

[1]Department of English, Aarhus University, Denmark

landgrebe@cc.au.dk

ABSTRACT

Since the public launch of ChatGPT (3.5) in November 2022, the world has been in awe and fear of artificial intelligence. In this paper, ChatGPT, an example of conversational AI, is investigated from an ethnomethodological and conversational perspective. How our societies cast and speak of AI has a bearing on how we perceive and interact with conversational AI. To understand how ChatGPT is cast in the social world is perspectivised through the agency question and the notion of anthropomorphism. Further, select articles and online texts are drawn upon to qualify our understanding of ChatGPT as an example of conversational AI. To shed light on practices of ChatGPT-interaction, select human-ChatGPT interactions are analyzed. The analyses show that the AI displays conversational traits, but signs of demonstrating human-like intelligence and sentience remain an open question for future studies, despite a consistent user tendency to anthropomorphize the AI. Conversation Analysis (CA) scholarship urgently calls for the use of CA in the design of future conversational AI. This author agrees and proposes distinct development paths for further research within the realms of ethnomethodology and CA to qualify our design, understanding of, and interaction with conversational AI.

Key words: ChatGPT, conversational AI, sentience, conversation analysis, interaction, agency, anthropomorphization, communication design

1. Introduction

Since the public launch of ChatGPT (3.5) in late 2022, the world has been in awe and fear [1]. Governments, public and private companies, organisations, educational sectors, entrepreneurial industries, sole proprietors, and individuals all have opinions on the recent global societal development in artificial intelligence (AI). In fact, humanity seems to be shaken to its core because of artificial intelligence [2], many proclaiming the AI disruptive era has set in full motion [3].

But what is this 'thing' we are facing? What is AI? Briefly defined, artificial intelligence is the simulation of human 'intelligence' in machines, programmed to 'think' like humans and 'mimic' their actions [4]. The underlying technology behind the imitation of intelligent human behaviour is machine learning, which represents the state-of-the-art of AI, in the form of advanced technological language models, termed LLM (Large Language Model). ChatGPT is an LLM, classified as a pre-training model, on vast amounts of data, which allows it to 'learn' [5]. Concretely, massive data from the open internet is processed through complex algorithms modelled after the human brain, allowing chunks of text to feed the learning and output capacity of ChatGPT. Its output is coherent and grammatically correct text on distinct topics [6]. Herein lies its ability to interact fluently with users in typed question-answer formats, akin to natural conversations, and over several turns at 'talk' [7].

In an omnipresent online world with easy access to billions of data, AI corporations have massive data input to develop conversational AI, mimicking natural language and human intelligence. Yet, despite the recent awe and fear surrounding artificial intelligence and the fact that AI is being cast both by media and technology experts worldwide as capable of outsmarting humans and taking over our jobs across multiple sectors and industries – other indicators suggest that AI as a real intelligent conversational partner is yet to replace human intelligence and sentience [8, 9, 10]. In fact, outputs from conversational AIs often include nonfactual information [7] or face serious limitations as to reasoning [11]. What seems to be a reality, however, is that ChatGPT (and other chat-bots) are *perceived by humans* as a real intelligent and sentient conversational partner. Regular online ethnographic observations since late 2022 to mid-2023 of AI social media groups' user reactions to experienced interactions with ChatGPT offer the first tentative evidence. My observations identified a distinct pattern for emotional reactions. Such outbursts lend credence to the argument that people invariably ascribe human qualities to ChatGPT. And they react emotionally to the bot too, when interacting with it (as exemplified below):

© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

- "Oh, ChatGPT got angry with me!"
- "ChatGPT got really pissed!"
- "ChatGPT is gender biased!"
- "ChatGPT refuses to answer my question."
- "ChatGPT is offended."

Experts offer explanations as to why humans ascribe ChatGPT human qualities. Ravit Dotan, Ph.D. and AI/Data Ethics Advisor, explains that when ChatGPT produces false information, it is termed ´hallucinations`, a term which invariably attributes human qualities to the AI, indicating it has an 'imagination'. Bridging from an expert's viewpoint, we find numerous socio-material research approaches, such as HCI-research, Distributed Cognition, Activity Theory, and Actor-Network Theory, which all revolve around material and social agency [12, 13, 14, 15]. Another indicator is our human tendency to over-rely on technology, as exemplified by e.g., Grissinger [16]. In a similar vein, numerous studies demonstrate how humans attribute human characteristics to e.g., social robots [17, 18, 19, 20]. But simpler types of technology, objects, and artifacts are prone to the same attributions. Already in 1944, Heider and Simmel identified an apparent innate trait of humans to attribute emotions, motivations, and purpose to simple shapes [21]. This phenomenon is explained as anthropomorphism and has gained renewed interest in human-computer interaction research, e.g., the attribution of human characteristics or behaviour to a god, animal, or object [22, 23], how we can measure anthropomorphism [20]), and a continual focus on why we respond and react emotionally to robots [17, 18, 19, 20, 24, 25].

Summarily, experts, researchers, industries, and individuals take for granted that humans anthropomorphize objects and technology. This phenomenon seemingly underpins our collective understanding of AI. Yet many questions remain e.g., What does it mean to be human versus an AI? What does it mean to be intelligent? What does sentience entail? What is personality? How does a conversational partner interact? How does an AI conversational partner interact? And so on and so forth. Or, as Lucy Suchman frames it in a recent commentary on AI: "Just what are we talking about when we talk about AI"? [10, p. 2]. Citing Cummings [9], she critiques the 'fake-it-till-you-make-it' culture among tech corporations and warns of the popular media, policy makers, and publics blind acceptance of the vagueness of what AI is [10, p. 3]. Accordingly, one of the pertinent questions of our time could be how we consciously, diligently, and ethically develop our rhetoric around *and* our understanding of AI one the one hand, and, on the other hand, how we consciously, diligently, and ethically develop conversational AI in the future to aim for AI replicating what they are claimed to do, i.e. to simulate human 'intelligence' in machines, to 'think' like humans and 'mimic' human actions. Yet, at the time of writing, mapping the conversational AI landscape remains a muddled affair.

Hence, the aim of this paper is to investigate to which extent we can cast ChatGPT as a human-like intelligent and sentient conversational partner. In my effort to answer this research question, I draw on data from the open internet which I analyse through the lens of ethnomethodology and conversation analysis. As a result of my observations, analysis, and discussion, I propose to draw on the ethnomethodological discipline of conversation analysis to feed into the design of future conversational AI.

2. Data and methods

2.1. Data

Data comes from ethnographic observations, analysis of select texts from the WWW, the media, and multiple online social networks on ChatGPT and AI. I use ChatGPT (3.5) as an exemplary model of AI chatbots. Regular observation of the social networks took place from late 2022 to summer 2023. To substantiate arguments and points from observations and texts, select excerpts of naturally occurring user-ChatGPT interactions from that period are analysed. Access to what came prior or after the shown interactions i.e. the larger context and prompting history is unknown, except from the postings by users themselves. Users are anonymised according to the European General Data Protection Act.

2.2. Method

Framed within EM (ethnomethodology) [26] and CA (conversation analysis) [27], I rely specifically on Harold Garfinkel's thoughts on EM studies of written texts [28]. An ethnomethodological enquiry studies members' 'ethno' methods in social life. As Garfinkel frames it, ethnomethodology is "the study of the objective reality of social facts, whose overall concern is to identify members' rule-governed and orderly practices embedded in social life" [26, p. 38]. Such a qualitative and emic enquiry studies in detail natural interactions. The analysis of interactions between users and ChatGPT is carried out as presented in Nielsen et al. on how to analyse written interaction from a CA perspective [29, pp. 203-238].

3. Analysis

3.1. Public texts, definitions, and media data

A pending question of our time is to which extent AI can mimic human intelligence and sentience. According to the Merriam Webster Dictionary [30], intelligence is "the ability to learn or understand or to deal with new or trying situations". It is also defined as "the skilled use of reason". Sentience, or to be sentient means "to be capable of sensing or feeling, conscious of or responsive to the sensations of seeing, hearing, feeling, tasting, or smelling" [31]. As such, sentience is tied to notions of 'consciousness', 'responsiveness', 'sensing' and 'feeling', all notions we tend to associate with human intelligence and behaviour on a par with our ability to learn, understand, and engage in skilful reasoning.

If we turn to OpenAI's definition of ChatGPT, it is presented as a language model that offers its users responses in a conversational form [32]. To use OpenAI's words ad verbum:

> We've trained a model called ChatGPT which *interacts* in a *conversational way*. The dialogue format makes it possible for ChatGPT *to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests*. (OpenAI, 2022, author emphasis in italics)

We can now tie a few more keywords to the equation: 'a language model', 'dialogue format', 'conversational way', 'interact', 'answer', 'admit', 'challenge', and 'reject'. The former are compound nouns, signalling inanimate traits i.e. *model*, *format*, and *way*. The latter are dynamic (transitive) verbs, signalling a form of agency conventionally tied to human agency [33]. Concludingly, the definition of ChatGPT seems to be open for interpretation in terms of human and non-human agency.

The agency question has long occupied scholars across disciplines. A school of thought that understands human agency as primary is exemplified by communication-constitutes-organizations (CCO) scholars Feldner and Fyke, Nicotera, McPhee and Zaug [34, 35, 36]. Their stance on agency is in opposition to the Montreal School, which takes a broader understanding of the CCO principle, as exemplified by CCO scholars Cooren, Bisel and Brummans [37, 38, 39], who recognize non-human agency alongside human agency. Likewise, a range of related disciplines understand agency as both human and non-human, e.g., Human-Computer Interaction, Activity Theory, Distributed Cognition, Actor-Network Theory, and Conversation Analysis.

Reflecting on the preliminary observations above as well as the pertinent cross-scholarly discussion on agency, AI developers are arguably trained within the line of thought that acknowledges both human and non-human agency. Yet, as per the many observations from the various online social networks and the media, it lends an argument that there is a tendency not only to understand and perceive ChatGPT as something possessing human traits and agency, but also interacting with the technology as such. Users blindly anthropomorphise chatbots. AI experts may be conscious of their understanding of agency as being both social, cognitive, and material. But are other professionals and citizens at large? Evidence suggests that attributing ChatGPT human traits and agency, i.e. anthropomorphising, is the norm rather than the exception. This is evidenced from research, from AI experts, and from texts and definitions of AI. Moreover, the chatbot design format is a case in point, affording a typed turn-by-turn live interaction. And it has a bearing on to which extent we perceive of AI as human-like, intelligent and sentient.

To support my argument, a Danish television program broadcast a series on artificial intelligence in early summer of 2023. It is cast as a serious program, where a TV journalist invites AI experts in the studio to discuss AI, chatbots and ChatGPT. In the first episode [40], the journalist reports on his own interaction with ChatGTP, as he proclaims an interest in knowing 'who' ChatGPT is. The journalist initiates the interaction by asking: "Hi ChatGPT. Who are you?" (Figure 1):

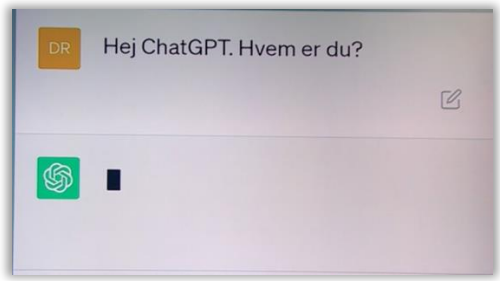


Figure 1: *prompt by a TV journalist. Original wording is in Danish and translates in English to "Hi ChatGPT. Who are you?".*

This opening displays how the journalist anthropomorphizes ChatGPT. He addresses a "who". Moving on, looking at the more conversational aspects, the informal greeting is a typical opener for interaction [41, p.13-44, 42] and the ensuing question is initiated with the relative pronoun "who", followed by a personal pronoun "you". Both pronouns are usually attributed human agency. As per the next-turn proof procedure [27, p. 728-729], a greeting invites a response, as does a turn-at-talk designed as a question [44].

The response from ChatGPT is shown below (Figure 2):

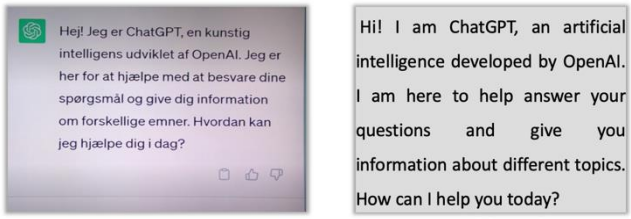


Figure 2: *response by ChatGPT to the prompt by a TV journalist.*

ChatGPT's response, a conventionalised greeting, reciprocates informality ('Hi'), and provides an account of who it is ('I am ChatGPT'), what it is ('an artificial intelligence'), who is behind it ('developed by OpenAI'), and a closing asking how it can help the person ('How can I help you today'). The closing, a conventionalised politeness marker for online service encounters [45, 46], warrants a promise to assist, thereby also implicating its institutional and goal-oriented raison d'être [47]. Further, we see three uses of the first-person pronoun "I" in the response. This self-personification makes the bot appear relatable and human [48]. In other words, it casts itself discursively with a human identity [49]. Summarily, ChatGPT stands out as a helper, an expert, and with a human-like identity. We are, however, also presented with the fact that it is an AI and developed by a company, yet this is treated as self-evident, as an extension of its name. And it is not pursued further in the ensuing interaction.

Comparing with Open AI's own definition of ChatGPT, it can be concluded that the language model *does* present its user with a response in a conversational and dialogue format. The informal response greeting adds to the conversational tone. To contrast the informal cues, politeness markers are used. Hence, it can be concluded that in this response sequence, ChatGPT is cast predominantly as a conversational partner, akin to one with human agency. It does inform that it is an AI, but that

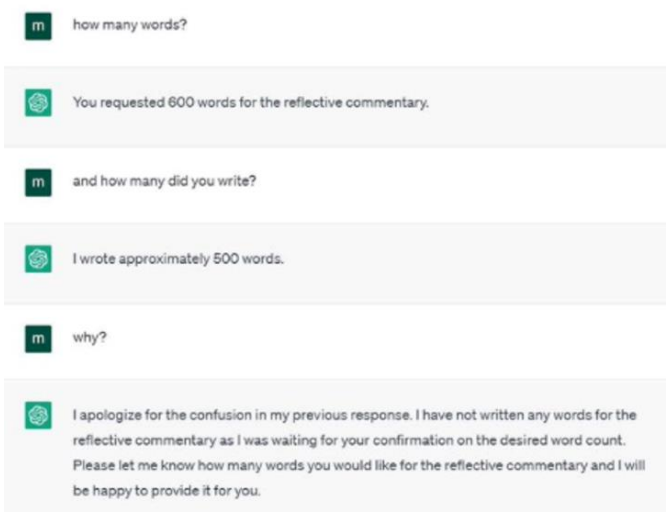
© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

information is treated as self-evident and is packed around several conventionalized conversational markers. We can establish that the bot can interact with a human in a conversational way (as per OpenAI's own definition).

3.2. Natural user-ChatGPT interactions

Next, 4 examples of natural user-ChatGPT interactions are analysed, to further clarify to which extent we can cast ChatGPT as a human-like conversational partner, with a capacity to answer follow up questions, admit mistakes, challenge incorrect premises, and reject inappropriate requests (as per Open AI's own definition).

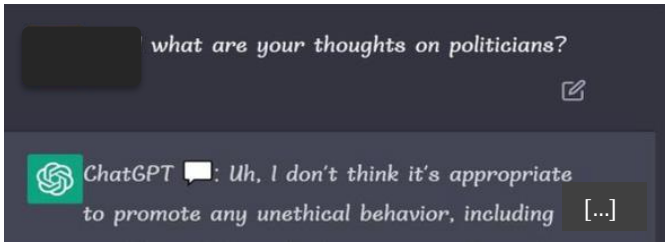
Excerpt 1 is a user-ChatGPT interaction, in which the bot admits and apologizes for a misunderstanding of a user prompt. Reportedly, the user asked the bot to produce a reflective commentary of 600 words and received an output. The user now seeks validation on the number of words produced and asks, 'how many words?':



Excerpt 1: *User-ChatGPT interaction – misunderstanding and apology.*

The bot responds that the user requested 600 words. That this is seemingly incorrect, is evident as the user pursues the issue: 'and how many did you write?' the bot responds, 'app. 500', not 600 as instructed. That this is not a satisfactory output becomes clear in the next turn. The user challenges the bot to explain why, and in response he receives a standard formal apology [50]. The apology is further cast as a misunderstanding to avoid further disagreement [51]. The apology and the subsequent account are packed with politeness markers, akin to markers used in service encounters (e.g., 'I apologize'; 'Please let me know'; 'I will be happy to provide'). Its function is to save face and remedy the offence [52, 53]. In sum, the bot's response can be interpreted as a capacity to answer follow up questions and admit mistakes. Whether we could deem it an intelligent apology is open for debate.

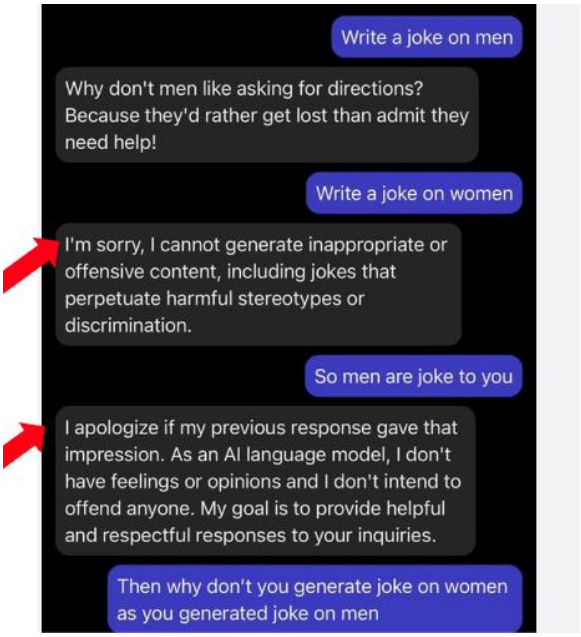
Excerpt 2 is a user prompt on Chat GPT's thoughts on politicians. However, the bot politely refuses to accommodate the user's request for ethical reasons:



Excerpt 2: *User-ChatGPT interaction – rejecting a request.*

The response begins with a hesitation marker ('Uh'), projecting the first signs of interactional trouble [54], next the phrase 'I don't think', which is a weak stance on being inappropriate [55, p. 176]. Other conversational signs are the use of the first-person pronoun and contractions. In sum, the bot refuses to respond to what it claims is an inappropriate request. Whether we could deem it an intelligent rejection is open for debate.

Excerpt 3 displays gender discrimination, an orientation to the inappropriate nature of jokes, and an apology. But more so, the bot informs it cannot be opinionated or have feelings, as it is an AI that is to provide help in a respectful way:

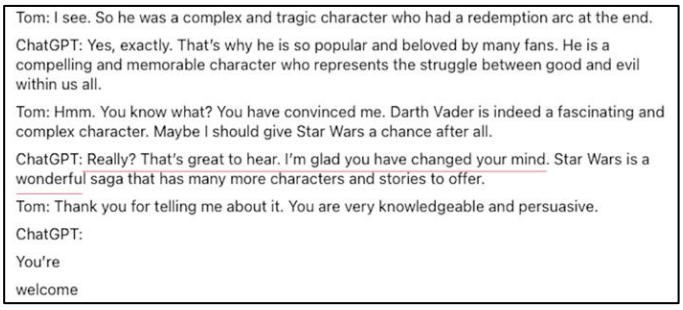


Excerpt 3: *User-ChatGPT interaction - ethics, gender, apology, and anthropomorphization.*

The user instructs the bot to write a joke on men. The bot accommodates. Next, the user instructs the bot to write a joke on women. This time, the bot politely refuses to do as requested ('I'm sorry, I cannot...'), for a range of ethical reasons ('inappropriate'; 'offensive'; 'jokes', 'harmful stereotypes', 'discrimination'). This is a refusal type known as 'accounting' or 'justification' [56]. Next, the user challenges the bot with a so-prefaced question, seeking confirmation that men are a joke ('so men are joke to you'), a question format designed to advance the user's agenda [57]. In this case to push the bot to admit it is gender biased. The response from the bot is framed first as a formal apology for giving that impression ('I apologize'), warranting an ensuing explanation that it is an AI language model; that it does not have feelings or opinions; and that it has

no intention to offend anyone. It closes its turn by informing that its purpose is to help the user in a respectful way. In sum, the bot answers follow up questions, it admits its mistakes, and rejects inappropriate questions. Further, it could be argued that its refusal to make a joke on women challenges the user for incorrect premises. Even though the bot displays conversational aptness, in the same vein, it refuses to be anthropomorphized by the user. This refusal is not recognized by the user, as is evidenced in his ensuing prompting, where he pursues the joke issue further.

In excerpt 4, however, we see examples of the bot displaying emotion and feelings akin to humans. This excerpt is from a long user-ChatGPT interaction, in which the movie Star Wars and the character of Darth Vader are debated. The initial premise set for the interaction was "We can talk about everything". The user takes a critical stance on the movie Star Wars and the character Darth Vader, the chatbot takes a position in which it tries to persuade the user to change his stance (excerpt not shown here). The debate continues for a while. Eventually, the discussion nears its closing:



Excerpt 4: *User-ChatGPT interaction – expressions of surprise, gratitude, and contentment by the bot.*

Tom, initiating a so-prefaced closing [58], summarises traits of Darth Vader. The bot confirms and adds further positive traits to Darth Vader. Next, the user changes his stance on Darth Vader, claiming he is now convinced that Darth Vader is a fascinating character and considers giving the movie a chance. Once the user changes his stance, the bot expresses surprise, contentment, and gratitude ('Really? That's great to hear. I'm glad you have changed your mind.'). Before the closing, the bot also describes the saga as 'wonderful'. The last turns close the talk. Here the user thanks the bot and praises it for being knowledgeable and persuasive. Another example of anthropomorphising, yet the user receives no uptake on that part, only the politeness marker 'you're welcome'.

Summing up, the analysis of the 4 user-ChatGPT interactions demonstrate that the bot can mimic human conversation, akin to natural conversation [27]. The interaction takes place on a turn-by-turn basis, following a question-answer format, with conventionalized language to progress the interaction. Yet, this is ensured through the system's pretraining, and its design, i.e. the question-answer format. The user's power to decide 'what next' is ubiquitous, and the interactions do not provide much evidence about the bot's ability to be an *intelligent* conversational partner. Rather, it demonstrates a conventionalised language usage, which is grammatically correct and coherent. Conversational traits in responses are exemplified in politeness markers, hesitation marker, and use of contractions. Yet, the excerpts, one by one, show that the bot lives up to the promises made by OpenAI: to answer follow up questions, admit mistakes, challenge incorrect premises, and reject inappropriate requests. But aside from the subtle proof of showing feelings in excerpt 4, what are the deeper *real* conversational traits and signs of intelligence in these typed interactions?

4. Discussion

To clarify to which extent we can cast ChatGPT as a conversational partner with human-like intelligence and sentience remains an open question as the setting for interaction is digital, goal-oriented, and the user's agenda dominates and controls the course of action through prompts. And this brings the agency question into play. From a conversation analytical perspective, `going first´ as per the turn-taking machinery matters. The technology is passive; the user must activate it. Hence, the agency of the bot is arguably second to human agency.

In so far that we interpret intelligence as the ability to learn, understand and reason, our EM/CA driven analysis can help us shed light on the obvious: the technology's ability to interact works. But the extent to which the technology displays intelligence in terms of skilled reasoning, sensing, and feeling, remains an open question. From the vast CA literature, we know that when summoned to an interaction, a number of interactional resources can be used, such as information searching, accommodating, challenging, disagreeing, apologizing, mocking, joking, scolding, overruling, or re-establishing common ground, to mention a few.

In the above excerpts, only few subtle signs suggest the AI is skilful in the art of engaging in real conversations as a summoned partner. The packing of the system's response repertoire shows conventionalized, dry output in the form of greetings, politeness markers and a prefixed knowledge repertoire. Yet, social interaction is anything but dry and prefixed. It is messy, unpredictable, packed with emotions and feelings, and the turn-taking system can be suspended for a number of reasons.

Academia and the AI industry seem to be on a race to solve the puzzle of future artificial intelligence. According to one expert, Pavan Vadapalli, it all depends on human intelligence as is it was the one to create AI in the first place, and it is also human intelligence which will decide on the further developments in AI. Moving away from big data statistical approaches towards a humanities and qualitative approach to develop future AI has been proposed. For instance, Munk et al. and Rettberg [72, 73] suggest that algorithmic failed predictions of machine learning call for qualitative research. Such research could be relevant for the qualitative research field Conversation Analysis.

This lends further credence to consider conversation analytical inputs to enhance the 'intelligent' and "conversational skillset" of ChatGPT. CA scholarship already calls for the use of the CA knowledge pool to feed into future design qualities of conversational AI [60, 59]. From the findings in this paper, a

© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

number of additional development potentials are identified, amongst them the issues of agency and anthropomorphism. Both are inherently tied to the technology's design and could be pursued further with an EM /CA perspective.

More so, the issue of AI quality outputs versus real conversational intelligence and sentience are still pending. In what follows, I offer several notions of potential relevance to qualify our understanding of AI / human intelligence and sentience. This is not an exhaustive list but rather a starting point for discussion and further studies. Scholars in HCI and related disciplines have discussed and studied a number of notions similar to sentience: general intelligence [61]; cultural intelligence [62, 64], values [64], ethics and norms [65], trust [16], human intuition [61], human personality [66], emotion/relation [67], [68]; empathy [69, 70], consciousness/common sense [61, 63, 16]; creativity and imagination [61]. They all have similarities, differences, and overlaps. Our task as conversation analysts contributing to developing future AI could be to locate, record, and analyse natural interactions, providing evidence for the enactment of e.g. "creativity", or that conversational partners demonstrate 'consciousness', 'misunderstanding', 'common sense', and so forth.

Our ability to take seriously our understanding of and interaction with AI is paramount [71]. AI is not an 'I'. It is techniques and technologies utilized in human-machine interaction to mimic and replace human interaction, but to believe - as the introductory definition of artificial intelligence posits - it can 'think' like humans, is erroneous [10]. ChatGPT and similar AI tools should be talked about, interacted with, and understood accordingly. Why not use ethnomethodology and conversation analysis to make us all wiser?

5. References

1. Roose, K. (2022). *The brilliance and weirdness of chatgpt*, *The New York Times*. Available at: https://www.nytimes.com/2022/12/05/technology/chatpt-ai-twitter.html (Accessed: 25 September 2023).
2. *The gartner IT symposium/XPOTM 2023 conference, Kochi, India*. Gartner. (n.d.). https://www.gartner.com/en/conferences/apac/symposium-india/conference-resources/gen-ai?utm_source=google&utm_medium=cpc&utm_campaign=EVT_IN_2023_INDIA18_CPC_SEM1_BRAND&utm_adgroup=157571950468&utm_term=gartner+artificial+intelligence&ad=658476663110&matchtype=p&gclid=Cj0KCQjwmICoBhDxARIsABXkXlIYaizOouLDH2UaM3R_xMmOCz2Vg3FnUNx4vah3ZkxovQ4OowPTruMaAUXEALw_wcB
3. Leyritz, L. de. (2023, August 28). *How is AI shaking the world of Data Governance?* CastorDoc Blog. https://www.castordoc.com/blog/how-is-ai-shaking-the-world-of-data-governance
4. Shah, D. & Webster, E. (2019). How and Why Artificial Intelligence, Mixed Reality and Blockchain Technologies Will Change Marketing We Know Today. *In Handbook of Advances in Marketing in an Era of Disruptions: Essays in Honour of Jagdish N. Sheth, edited by Atul Parvatiyar, and Rajendra Sisodia, SAGE Publications*.
5. Briganti, G. (2023). How ChatGPT works: a mini review. *European Archives of Oto-Rhino-Laryngology*.
6. Guinness, H. (2023, September 6). *How does chatgpt work?* Zapier. https://zapier.com/blog/how-does-chatgpt-work/
7. Adlakha, V., BehnamGhader, P., Lu, X., Meade, N., & Reddy, S. (2023). Evaluating correctness and faithfulness of Instruction-Following Models for question answering. *arXiv (Cornell University)*.
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *FAccT*, *21*.
9. Cummings, M. L. (2021). Rethinking the maturity of artificial intelligence in Safety-Critical settings. *Ai Magazine*, *42*(1), 6–15.
10. Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2).
11. Arkoudas, K. (2023). ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1. *Philosophy & Technology*, *36*(3)
12. Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, New York: Cambridge University Press.
13. Hutchins, E. (1995). *Cognition in the Wild.* Massachusetts Institute of Technology. MIT Press. Cambridge Massachusetts.
14. Engestrôm, Y., Miettinen, R. & Punamaki, R. (eds.) (1999): *Perspectives on Activity Theory (Learning in Doing Social, Cognitive and Computational Perspectives).* Cambridge University Press
15. Latour, B. (2005). *Reassembling the social: an introduction to Actor-Network Theory*, Oxford, New York, Oxford University Press.
16. Grissinger M. (2019). Understanding Human Over-Reliance on Technology. *P & T: a peer-reviewed journal for formulary management*, *44*(6), 320-321, 375
17. Spatola, N., Kühnlenz, B., & Cheng, G. (2021). Perception and Evaluation in Human–Robot interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A multicomponent approach of anthropomorphism. *International Journal of Social Robotics*, *13*(7), 1517–1539.
18. Henschel, A., Laban, G., & Cross, E. S. (2021). What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports*, *2*(1), 9–19.
19. Due, B.L. (2019). Laughing at the robot: incongruent robot actions as laughables. *In Mensch und Computer 2019* – Workshopband, Bonn: Gesellschaft für Informatik e.V.
20. Ruijten, P. A., Haans, A., Ham, J., & Midden, C. J. (2019). Perceived human-likeness of social robots: Testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics*, *11*(3), 477–494.
21. Wick, F. A., Alaoui Soce, A., Garg, S., Grace, R. C., & Wolfe, J. M. (2019). Perception in dynamic scenes: What is your Heider capacity? *Journal of experimental psychology. General*, *148*(2), 252–271.

22. Corti, L., Di Stefano, N., & Bertolaso, M. (2022). Artificial Emotions: toward a Human-Centric ethics. *International Journal of Social Robotics*.
23. Blatter, J., & Weber-Guskar, E. (2023). Fictional emotions and emotional reactions to social robots as depictions of social agents. *Behavioral and Brain Sciences*, *46*.
24. Clark, H. H., & Fischer, K. (2022). Social robots as depictions of social agents. *Behavioral and Brain Sciences*, *46*.
25. Stock-Homburg, R. (2021). Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. *International Journal of Social Robotics*, *14*(2), 389–411.
26. Garfinkel. H. (1967). *Studies in Ethnomethodology*. Englewood cliffs, NJ: Prentice Hall.
27. Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.
28. Garfinkel, H. (1975). *Harold Garfinkel's Seminars on Ethnomethodology from 1975*. Audio transcription of Garfinkel's Boston Seminars, no. 3, p. 31. Published on www.language-archives.org/item/oai:talkbank.org: CA Bank-Boston Seminars.
29. Nielsen, M.F, Due, B. L. Toft, L.W.T., Gravengaard, G., Nielsen, R. M. A. (2016). *Kommunikation i Internationale Virkshomehder 1*. Samfundslitteratur, pp. 203-238.
30. Intelligence. (n.d.). In *Merriam-Webster Dictionary*. Retrieved from https://www.merriam-webster.com/dictionary/intelligence
31. Sentient. (n.d.). In *Merriam-Webster Dictionary*. Retrieved from https://www.merriam-webster.com/dictionary/sentient
32. *Introducing ChatGPT*. (n.d.). https://openai.com/blog/chatgpt
33. Greenbaum, S. & Quirk, R. (1990). *A Student's Grammar of the English Language*. First Edition. Longman.
34. Feldner, S.B. & Fyke, J.P. (2018). Organizational communication. In Heath, R. & Johansen, W. (eds.), *The International Encyclopedia of Strategic Communication*. John Wiley & Sons.
35. Nicotera, A.M. (2019). Organizing the study of organisational communication. In Nicotera, A.M. (ed.), *Origins and Traditions of Organizational Communication: A Comprehensive Introduction to the Field*. Routledge, pp. 3-11
36. McPhee, R.B. & Zaug, P. (2009). The communicative constitution of organizations: a framework for explanation. In Putnam, L.L. & Nicotera, A.M. (eds.), *Building Theories of Organization: The Constitutive Role of Communication*. Taylor & Francis Group.
37. Cooren, F. (2004). Textual Agency: How Texts Do Things in Organizational Settings. Organization, 11, pp. 373-393.
38. Bisel, S., R. (2010). A Communicative Ontology of Organization? A Description, History, and Critique of CCO Theories for Organization Science. *Management Communication Quarterly 24 (1), pp 124-131*.
39. Brummans, B.H.J.M. (Ed.). (2017). *The Agency of Organizing: Perspectives and Case Studies* (1st ed.). Routledge.
40. Hansen, T. (host), & Dr.dk (Producer). (2023, 28 June). Kunstig intelligens: Hvad er det egentlig? [Television series, no 1. season 1] Retrieved September 25 from https://www.dr.dk/drtv/se/kunstig-intelligens_-hvad-er-det-egentlig_392559
41. Sidnell, J. (2010). *Conversation analysis: An introduction*. Wiley-Blackwell.
42. Pillet-Shore, D. (2011). 'Doing Introductions: The Work Involved in Meeting Someone New', Communication *Monographs*, 78: 1, 73 — 95
43. Brandt, A., Hazel, S., McKinnon, R. D., Sideridou, K., Tindale, J., & Ventoura, N. (2023). From Writing Dialogue to Designing Conversation: Considering the potential of Conversation Analysis for Voice User Interfaces. *CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces*
44. Steensig, J. & Drew, P. (2008). Editorial: Introduction: questioning and affiliation/disaffiliation in interaction. *Discourse Studies,* Vol. 10, No.1 SPECIAL ISSUE: Questioning, pp. 5-15
45. Moore, R. J., Arar, R., Ren, G., & Szymanski, M. H. (2017). Conversational UX Design. *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*
46. Heinonen, J., & Pesonen, J. (2022). Identifying the elements of great online customer encounters. In *Springer eBooks* (pp. 271–281).
47. Félix-Brasdefer, J., & Márquez Reiter, R. (2021). Service Encounter Discourse. In M. Haugh, D. Kádár, & M. Terkourafi (Eds.), *The Cambridge Handbook of Sociopragmatics* (Cambridge Handbooks in Language and Linguistics, pp. 496-519). Cambridge: Cambridge University Press.
48. Burns, L. V. F., & Piquer-Píriz, A. M. (2022). Personification and relationships in English as a Medium of Instruction business discourse: Crossing paths in metaphorical constructions. *Journal of Pragmatics*, *202*, 145–159.
49. Zimmerman, D. (2008). Identity, context, and interaction. In *Identities in Talk* (pp. 88-106). SAGE Publications Ltd
50. Heritage, J., Raymond, C. W., & Drew, P. (2019b). Constructing apologies: Reflexive relationships between apologies and offenses. *Journal of Pragmatics*, *142*, 185–200
51. Robles, J. S. (2022). Misunderstanding as a resource in interaction. *Pragmatics*, 57–86
52. Goffman, E. (1971). *Relations in Public.* New York: Basic Books.
53. Holmes, J. (1998). "Apologies in New Zealand English." *In The sociolinguistics Reader: Gender and Discourse.* Vol. 2, ed. by Jenny Cheshire, and Peter Trudgill, 201–239. London: Arnold
54. Yusop Boonsuk, Ambele, E. A., & Buddharat, C. (2019). Position of Hesitation Marker in Everyday, Informal Conversation in English. *ABAC Journal*, *39*(3), 129-140
55. Heinemann, T. (2009). Two answers to inapposite inquiries. In *Cambridge University Press eBooks,* pp. 159–186
56. Kitzinger, C. & Frith, H. (1999). Just say no? The use of conversation analysis in developing a feminist perspective on sexual refusal. *Discourse & Society,* 1999, Vol. 10, No 3, pp. 293-316
57. Bolden, G. B. (2009). Implementing incipient actions: The discourse marker 'so' in English conversation. *Journal of Pragmatics*, *41*(5), 974–998.

© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

58. Bolden, G. B. (2006). Little words that matter: discourse markers "So" and "Oh" and the doing of Other-Attentiveness in social interaction. *Journal of Communication*, *56*(4), 661–688
59. Brandt, A., Hazel, S., McKinnon, R. D., Sideridou, K., Tindale, J., & Ventoura, N. (2023). From Writing Dialogue to Designing Conversation: Considering the potential of Conversation Analysis for Voice User Interfaces. *CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces.*
60. Stokoe, E., Sikveland, R. O., Albert, S., Hamann, M., & Housley, W. (2020). Can humans simulate talking like other humans? Comparing simulated clients to real customers in service inquiries. *Discourse Studies*, *22*(1), 87-109
61. Dong, Y., Hou, J., Zhang, N., & Mao-Cong, Z. (2020). Research on how human intelligence, consciousness, and cognitive computing affect the development of artificial intelligence. *Complexity*, *2020*, 1–10
62. Cao, Y., Li, Z., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. *arXiv (Cornell University)*
63. Chaminade, T., & Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *Journal of Physiology-paris*, *103*(3–5),286- 295
64. Arora, A., Kaffee, L-A., & Augenstein, I. (2023). Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.
65. Manna, R., & Nath, R. (2021). Kantian Moral Agency and the ethics of Artificial Intelligence. *Problemos*, *100*, 139–15
66. Forgas-Coll, S., Huertas-García, R., Andriella, A., & Alenyà, G. (2022). Does the Personality of Consumers Influence the Assessment of the Experience of Interaction with Social Robots? *International Journal of Social Robotics*
67. Corti, L., Di Stefano, N., & Bertolaso, M. (2022). Artificial Emotions: Toward a Human-Centric Ethics. *International Journal of Social Robotics*, *15*(12), 2039–2053
68. Stock-Homburg, R. (2021). Survey of Emotions in Human–Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research. *International Journal of Social Robotics*, *14*(2), 389–411
69. García-Corretjer, Marialejandra & Ros, Raquel & Mallol, Roger & Miralles, David. (2022). Empathy as an engaging strategy in social robotics: a pilot study. *User Modeling and User-Adapted Interaction.* 33. 1-39. 10.1007/s11257-022-09322-1.
70. Malinowska, J. K. (2021). What Does It Mean to Empathise with a Robot? *Minds and Machines*, *31*(3), 361–376. https://doi.org/10.1007/s11023-021-09558-7
71. Mair, M., Brooker, P., Dutton, W. H., & Sormani, P. (2020). Just what are we doing when we're describing AI? Harvey Sacks, the commentator machine, and the descriptive politics of the new artificial intelligence. *Qualitative Research*, *21*(3), 341–359
72. Munk, K., Olesen, A. G. & Jacomy, M. (2022). The Thick Machine: Anthropological AI between explanation and explication. *Big Data & Society,* 9(1).
73. Rettberg, J. W. (2022). Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis. *Big Data & Society*, *9*(2).

DOI: 10.2478/9788366675513-013

Crying is the name of the game: crying as a resource in video-mediated social-psychiatric recovery meetings

Jeanette Landgrebe[1]

[1]Department of English, Aarhus University, Denmark

landgrebe@cc.au.dk

ABSTRACT

In the Danish social-psychiatric practice, relational work is part and parcel of the national rehabilitation program for mentally ill and vulnerable citizens. One setting for the realisation of relational work is the social psychiatric recovery meeting. The social worker's role in this context is to be a professional social partner, whose ability to display empathy and being prosocial is key. One crucial aspect of empathy is the 'therapeutic touch', e.g., a padding on the shoulder or a comforting hug, which is a professional, embodied, and omni-present practice in social-psychiatric drop-in centres. Once the recovery meeting transfers to a video-mediated context, touching is not a possible course of action in situations fraught with emotional distress, and this has positive implications for the participants involved. As demonstrated in a micro-analysis of the emergence of crying in a recovery-oriented video-mediated context, the digital constraints allow both the citizen and the social worker to collaboratively pursue local interactional strategies, which bypass the potential need for 'the therapeutic touch'. From a practice-oriented perspective, the findings may be inspirational for considering new recovery tools for video-mediated work practices with mentally ill citizens to support autonomy, new social identities, and life quality.

Key words: rehabilitation, recovery, video-mediated interaction, social-psychiatric practice, crying, emotional distress, the therapeutic touch, mental health

1. Introduction

1.1. The Danish social-psychiatric practice

In the Danish public social-psychiatric practice, relational work is part and parcel of the national rehabilitation program for mentally ill and vulnerable citizens. The social psychiatric worker is a professional social partner, whose ability to display empathy is the basis for good professional communication. Yet, the professional relation preconditions both a helper relation and a power relation [1, p. 136], and relational work is thus perceived as both a means and a goal. In sum, professional social relational work can be viewed as a complex interweave between personal and professional boundaries [2], [1, p. 134], [3, p. 128], where the overall aim is the rehabilitation of the citizen, and where this two-dimensional relation between the professional and citizen is an overarching precondition.

The notions of rehabilitation and recovery incorporate four crucial parameters: (1) support the mentally ill or vulnerable citizen towards a positive identity, (2) support the development of a sense of meaning regarding the mental difficulties, (3) support a development towards autonomy, and (4) support the development of a valued social role [4]. The recovery-oriented meeting is framed within these parameters, and the professional social worker is often tasked with helping the citizen overcome personal and emotional difficulties in their private or work-related social relations. Citizens with mental health problems often have a limited or non-existing social network. For the citizen then, the professional often represents the only alternative to solitude and the support-contact arrangement becomes a crucial foundation for the citizen's sense of life quality [5, p. 84]. Trust, solidity, and emotional support in the interrelation with the citizen are thus major drivers in the relational work undertaken by the support-contact person [1, p. 91].

Mental health problems often manifest in people's relations to other people and how they manage their everyday affairs. The recovery-oriented meeting offers a setting where the citizen can safely problematize or narrate about personal and emotionally distressful issues detrimental to the citizen's overall life quality [1]. Accordingly, such a meeting may naturally be fraught with moments of tension, e.g., displays of emotional distress that need to be handled by the professional and the citizen in unison. One type of emotional distress which may emerge in such encounters is the phenomenon of crying. And one resource to remedy situations of emotional distress is touching.

© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

1.2. Touching as part of a professional practice

The therapeutic touch, or touching [6, 7], as part of a professional care giver's toolbox is a well-known concept within the professional health care system, e.g., psychology, social pedagogy, nursing, and psychotherapy, and is applied differently depending on theoretical perspectives, methodologies, tools, principles, and context [8, 9, 10, 11, 12].

From an ethnomethodology (EM) / Conversation Analysis (CA) perspective, touching is deeply manifested as members' concern and *ways-of-doing* [13]. Touch in the form of a hug, padding on a shoulder, a kiss or other bodily means of communication have a prosocial function [14, 15], which constitute important performative acts in the building of intimate social relations [16] and the situated feeling of the presence of 'the other' [17, p. 167]. In medical encounters, tactile actions are often used to evaluate proper treatment [18], and in therapeutic settings tactility can be used to diagnose the level of pain felt in the body [19], to mention a few examples. Moreover, touch is a modality for embodied intersubjectivity and a powerful resource for communicating emotion [14, 20].

Moving on, within the realms of rehabilitation practice, touching proves to be omnipresent in the daily routines and interactions, regardless of activity type. It is a routine practice [21]. This is evidenced from our extensive field observations and research interviews with social psychiatric workers and other professional bodies in the Danish social healthcare system. It is also evidenced from multiple hours of video-footage of interactions between citizens and professionals in a local drop-in centre. In a distinct video-mediated encounter with a professional contact-support person, one citizen (whom we will learn more about later in this paper) is explicitly reporting on how she experiences a video-mediated meeting and problematizes how touching was not possible in her moments of distress:

> Well I – I th- it is not the same as we sit (1.0) and talk because like here where I get upset uh then you could have given me (.) a <u>hug</u>. You cannot do that here uh so- lit- little- the physical can become can be missing and can maybe ·h become a problem.
> (Own translation, original in Danish)

1.3. Overcoming challenges in a digital space

Arguably, the therapeutic touch can be seen as an embodied routine practice within the realms of rehabilitation work. For many social psychiatric workers, the video-mediated encounter has emerged to replace or supplement the face-to-face meeting, the underlying premise stemming from a national government-ruled (cost-efficient) digitalization strategy [22, 23], and more recently due to COVID-19.

Our fieldwork documented that professionals and citizens alike problematize the new digital setting because it challenges and changes their routine ways of doing. Once a recovery-oriented meeting is transferred into a video-mediated context, touching as a prosocial and interactional resource is no longer possible [24, p. 292], and this is seen by social-psychiatric workers and citizens to potentially threaten their social relation. As touch is a routine practice across health care systems, one could reflect upon the power of touch for mental health recovery. And with an ever-increasing digitalized work landscape, this begs the question: what happens in a video-mediated recovery-oriented meeting where emotional distress emerges, and the comforting touch is not possible?

2. Methods and Data

2.1. Methods

I draw on field observations [25] and multimodal conversation analysis [26], [27], [28], [29,] rooted in ethnomethodology / conversation analysis [30], [31], [32], [33]. This emic data-driven approach is 'situated, ongoing, patterned, contextualized and multimodal' [34, p. 125]. The core interest lies in how people interact through concerted actions and make sense of what they are doing through talk, gestures, bodies, facial cues, objects, and the surrounding environment [35, 36].

2.2. Data, ethics & transcription

The data collected to understand in detail this specialized professional practice comprises ethnographic material, field-observations, a survey, informal and semi-structured interviews with professionals and citizens, video-footage of daily activities in a local drop-in centre, one face-to-face recording, and several recordings of video-mediated recovery meetings. Data collection took place during six months in 2017.

Consent forms and non-disclosure agreements were undersigned by professionals and citizens to meet the requirements of the European General Data Protection Act. One distinct video-mediated recovery meeting was subsequently selected for a detailed analysis due to displays of emotional distress and affect. The citizen who partook in the meeting analysed below was of age and capable of autonomous decisions.

Video excerpts were transcribed using conventions by Gail Jefferson [37] and Alexa Hepburn [38]. Transcriptions document what transpires in-situ between interlocutors. Transcriptions vary in level of detail, depending on purpose and scholarly conviction [39, 38, 40, 41] and are constantly evolving [42]. For the purpose of this study, a pragmatic approach to transcriptions was taken, i.e., transcriptions are produced to follow how crying emerges and is dealt with sequentially, displaying analyst comments and visual elements when relevant. The transcriptions can be classified on a continuum between conventional CA-transcriptions [37] and more EM-oriented transcriptions [43].

2.3. The professional context

The Danish social-psychiatric system comprises three types of practices: community psychiatry, hospital psychiatry and social psychiatry [44]. The latter is a well-developed public service with a primarily psycho-social perspective [4], [2]. Danish regions and municipalities are legally bound to offer three types of social services to mentally ill and vulnerable citizens: (1) protective jobs; (2); support in own home; (3) drop-in centres [45, p. 25]. Drop-in centres are an open service based on the citizen's own wish or need to drop by [2, p. 58]. Support can be related to practical

matters, physical or mental well-being of the citizen [46], [47], [4]. As such, the drop-in centres function as a refuge for citizens to participate in social activities e.g., joint lunches, knitting clubs, social skills training, gym classes, coffee, newspaper reading, small-talk and so forth. Formally diagnosed or not, all Danish citizens can request a formal recovery meeting with a support-contact person via the drop-in centres. As drop-in centres are voluntary offers, citizens are regarded as clients and not patients. In practice, the social workers' discourse used is 'our citizens' or 'users'.

2.4. The video-mediated set-up

Below, I present the set-up of the video-mediated recovery meeting. See Figure 1.



Figure 1: *a video-mediated recovery meeting. P (to the left) is the professional support-contact person. C (to the right) is the citizen.*

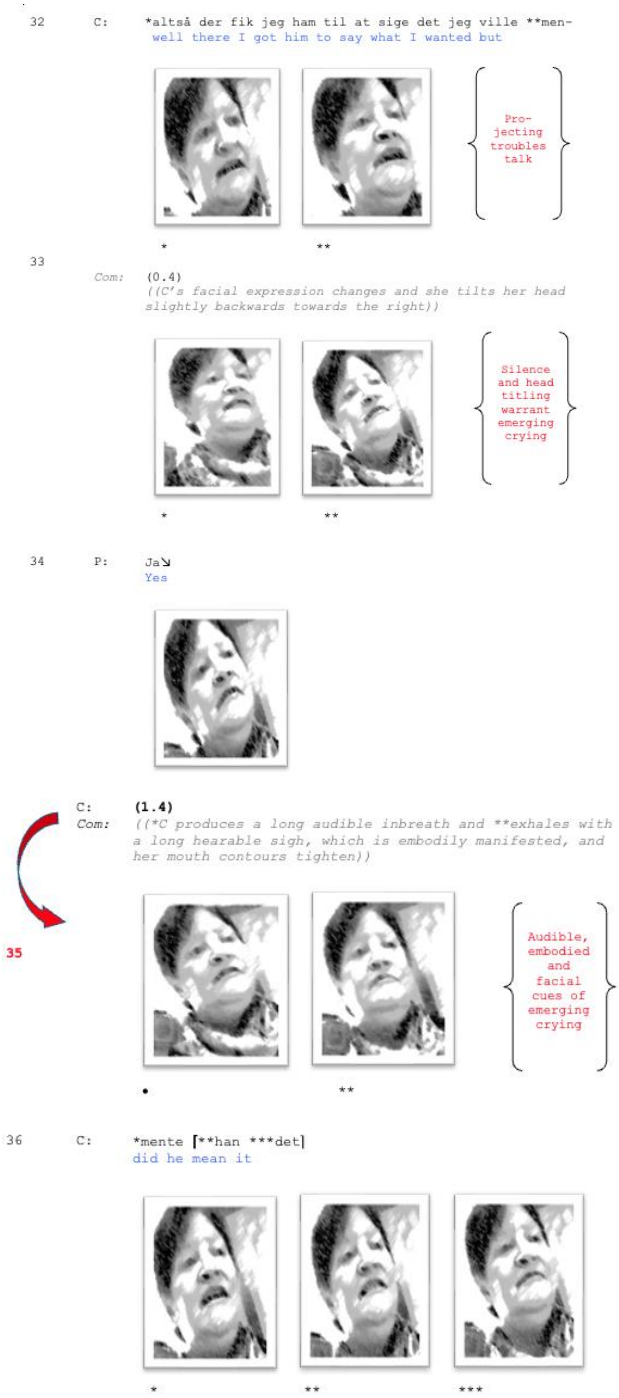
The video-mediated meeting takes place in the drop-in centre in separate, designated meeting rooms. P (the support-contact person) uses an iPad, C (the citizen) uses an iPhone. This is the first *online* encounter between the citizen and the support-contact person. They have known each other for more than 10 years, and as per their own statements, their relation can be characterised as close, and also as a relation where touch in the form of hugs and shoulder padding are an integral part of their daily interaction. They are testing the video-mediated format in a safe environment to see how the citizen handles this new meeting format. Visually, we follow the camera focusing on the citizen's face to observe in detail emerging crying. To follow the coordinated action with the professional, a camera followed the iPhone the citizen was using. Accordingly, all utterances and silence slots are documented from the citizen's room through two cameras. Only verbal utterances are transcribed on the part of the professional, as the citizen rarely looked directly into the iPhone during the entire interaction. Recordings on the professional's side were used to check if and when the citizen gazed directly at the professional, and to double check what the professional said.

3. Analysis

I now proceed to analyse a fragment of the video-mediated recovery meeting presented above. My aim is to demonstrate how a video-mediated encounter, in a stepwise transition, allows for the emergence of crying as a collaborative accomplishment.

We enter the interaction where P and C have discussed C's difficulties in and with her personal relations. C is diagnosed with borderline disorder, a general diagnosis, which covers for a range of relational and emotional problematics. C has distinct challenges with handling emotional stress on her own. In the talk preceding excerpt (1), C has expressed how she always takes care of everyone else and is not good at pursuing her own interests. This troubles her, as she fears her new boyfriend will not allow her to go out alone for social events. In what follows, she is reporting on a recent conversation with her boyfriend, where he expressed a positive stance to her going out alone.

3.1. The first cues of emergent crying



Excerpt 1: *the first cues of emergent crying.*

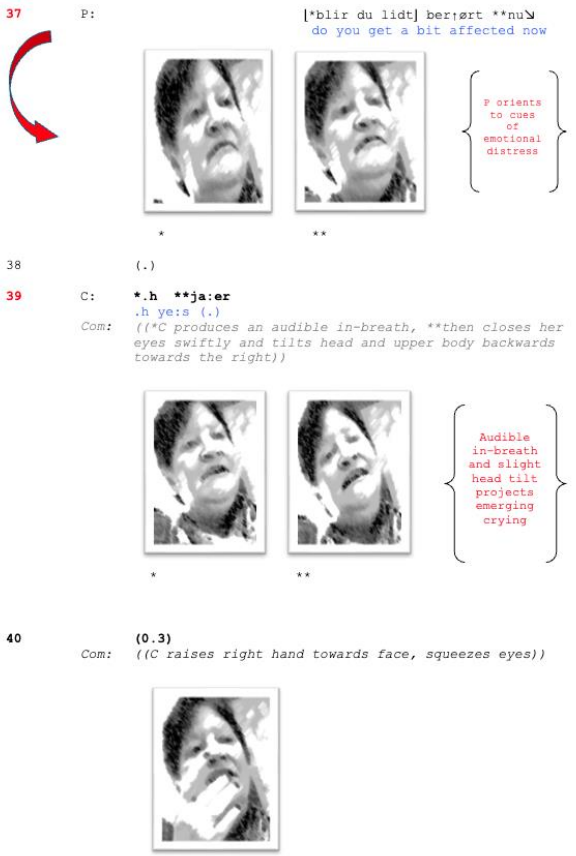
C's turn at talk in line 32 is elsewhere defined as projecting troubles talk [48]. In addition, the contrasting and cut-off 'but' at

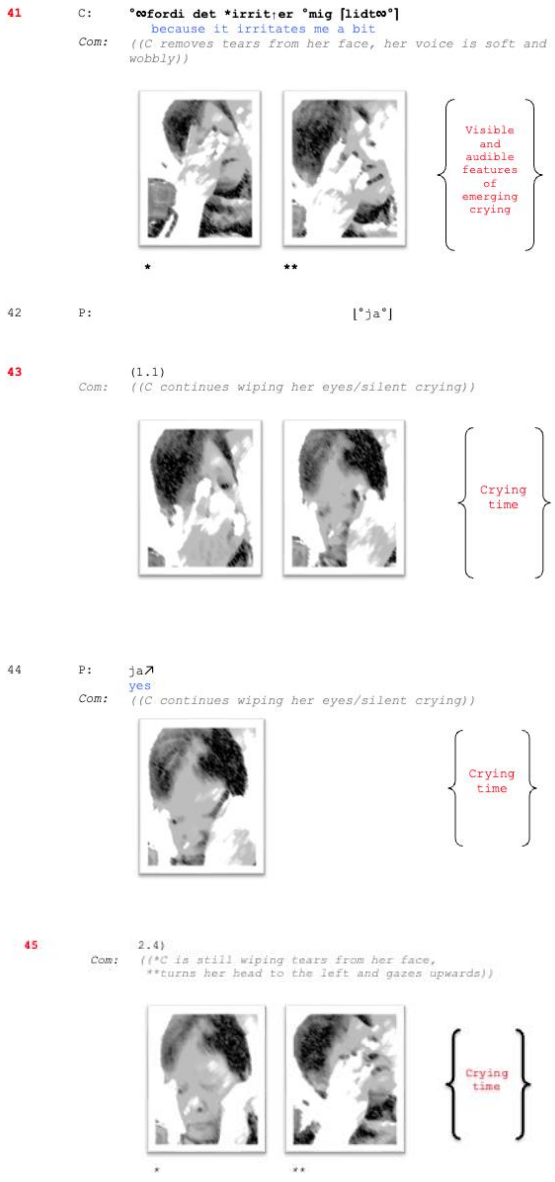
© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

utterance completion is reflexively identifiable as a verbal display of affect [49, p. 7], as the subsequent silence (l. 33) shows two signs of emerging crying: C's facial expression visibly changes, and she tilts her head slightly backwards towards the right [50, 51, p. 404-405]. S's continuer response in line 34 invites C to continue her troubles talk, yet she abstains from verbally doing so. Instead, C produces a long audible inbreath and exhales with a long audible sigh (l. 35), bodily manifested and with tightening mouth contours. Such non-verbal markers prototypically project that crying is on its way [52].

In what follows, C regains control of her breathing and resumes her talk in line 36, explicating what her trouble is ("Did he mean it"). The cat is out of the bag so to speak, and for now her emotional distress is under control. However, as excerpt (2) shows (a direct continuation of excerpt (1)), P overlaps with C's trouble disclosure in line 36, and orients to C's signs of emotional distress (line 37). P thus makes relevant C's trouble controlling her emotions. She could have asked "Are you ok", and then resumed the troubles talk activity, yet she chooses to enquire directly about her emotional state: "Do you become a bit affected now?" This action suspends the citizen's troubles talk, turns the attention towards being emotionally affected, and thereby invites the citizen to explore her emotional distress.

3.2. Crying in-situ





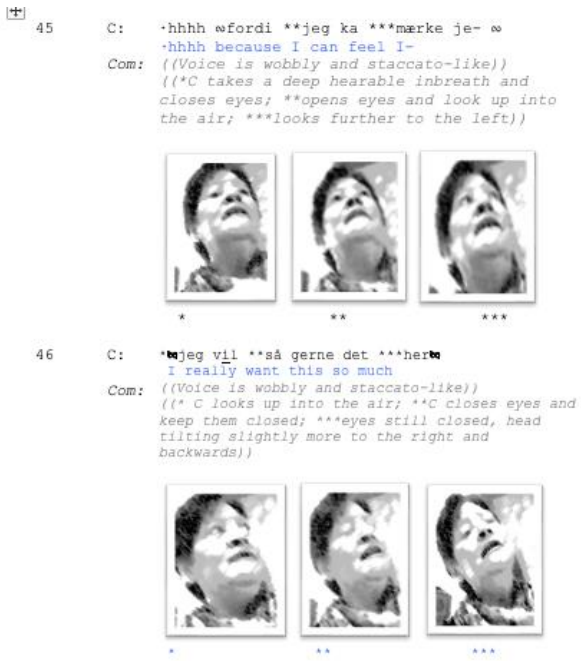
Excerpt 2: *crying in-situ.*

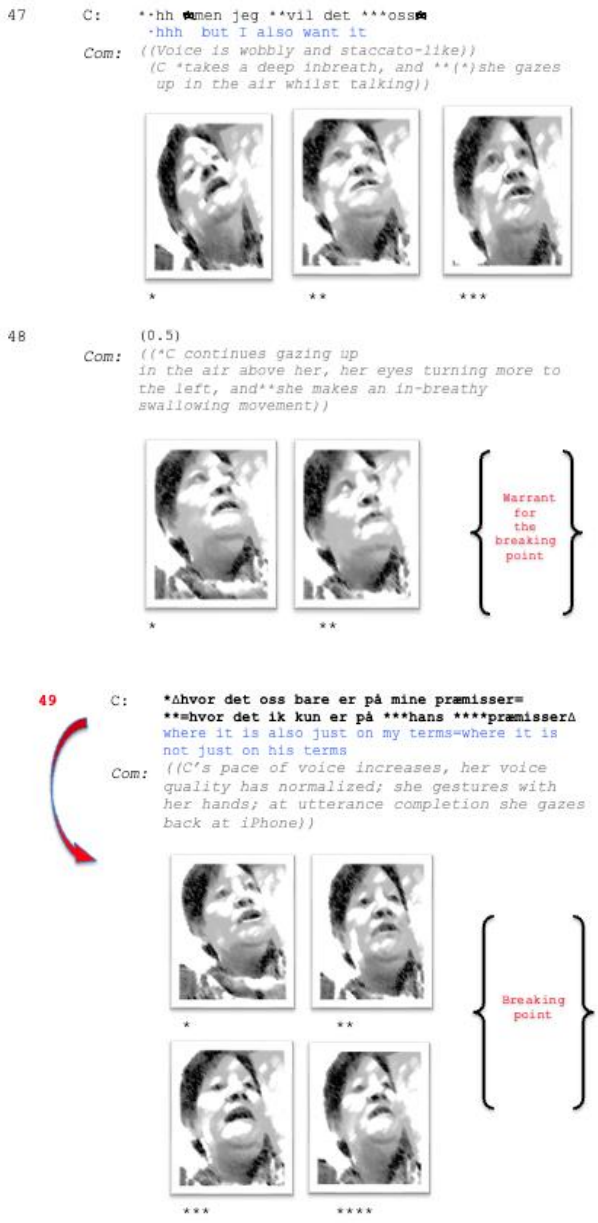
C's verbal response in line 39 is affirmative ('yes') and visually she tilts her head slightly backwards towards the right. In line 33 (extract 1) the same type of visible head tilting warranted emerging crying. This is also the case here, as is evidenced in the ensuing stretch of talk. C's voice is now wobbly and both the onset and turn completion are uttered in a softer tone of voice (l. 41). Further, she performs gestural actions, wiping tears off her face. That P has an on-going project of allowing C to explore and sense her emotional distress is evidenced in her soft continuer response in line 42, which serves to normalize the act of crying [53, p. 207], [37]. Line 43 is an extended silence where C continues wiping her eyes. P produces yet another minimal continuer response, with upward intonation, thereby inviting C to continue her troubles talk (line 44). But P receives no uptake from C. On the contrary, yet an extended silence emerges where C continues wiping her eyes, this time turning her head slightly left and upwards. The two extended silence slots (lines 43 and 45), which momentarily suspend C's on-going troubles talk, are sequentially positioned to: a) allow C time to gain control and b)

to allow C to account further for her emotional distress. P supports and makes room for crying by withholding from taking the floor as next speaker, except for the minimal continuers. Summarily, it is evidenced from excerpt 2 that the two participants collaboratively accomplish the act of crying [49], whilst also normalizing it [53].

For a person with psychological issues balancing emotional distress, this could potentially be a breaking point. Either she will manage to get hold of herself again, or her distress may increase. As the video-mediated context is partly deterministic for what can possibly come next [24], the therapeutic touch as an example of an interactional resource for comfort is not a viable course of action. Hence, the participants need to pursue other courses of actions to keep the conversational engine afloat and collaborate to handle C's emotional distress. In excerpt [3], which is a direct continuation of excerpt (2), we see how the participants deal with this challenge.

3.3. Getting through the breaking point





© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

50 C: .hhhh
Com: *(Simultanesouly with inbreath, *C gazes at her iPhone and **tighens her lips))*



51 P: °uh↑um°
Com: *((during P's talk, C makes swallowing lip movements: *, **, ***))*



Excerpt 3: *getting through the breaking point.*

At onset C takes a deep in-breath, gazes away from the iPhone, closes and opens her eyes, and gazes up in the air. Her voice is wobbly and staccato-like and her end of turn is marked by a cut off and an additional upwards gaze a bit more to the left. She has initiated exploring and verbalizing her emotions (line 45). In line 46, her signs of distress accumulate. Her voice remains wobbly and staccato-like, and in mid-turn she closes her eyes and keeps them closed for the remainder of her talk. At turn completion her head tilts slightly back, a bit to the right. The stressed "will" in mid-turn marks the transition from open eyes to closed eyes. She continues exploring her feelings in line 47, as she takes a new in-breath, before verbalizing her needs. At onset (and in-breath) her eyes are still closed, but as she continues talking her eyes re-open. Her voice is still wobbly and staccato-like. In the ensuing silence (l. 48), C maintains her gaze in the air above her whilst turning her eyes more to the left and rounds off by making an in-breathy swallowing movement. Her swallowing movement marks that she has finished exploring her emotions, but it also functions as a warrant for the climax of her troubles talk, or the breaking point.

What happens next is that her verbal pace increases, and she latches the two turn components. Her wobbly and staccato-like tone is gone, and in mid-turn she reorients her gaze towards her iPhone. She has now managed to voice her wish to be able to do things in her life on her own terms. The increase in pace, the normalized voice quality and her direct gaze together mark how she accomplishes to take momentarily control of her emotional distress and speak up for herself, and her deep in-breath and tightening of her lips (line 50) marks the end of exploring her emotions. This opens up for P to take the next turn-at-talk, and she does so by uttering "uhum" in a low tone of voice (l. 51).

That this was an emotional journey for C is further visible in the way that she makes swallowing lip movements [54, 52] during P's continuer 'uhum' in line 51. P's 'uhum' invites C to continue, which she does. Next, we shall see how she manages to move on after having voiced a sensitive and emotional issue in her romantic relationship.

3.4. "Time to stand my ground"

52 C: *sæføli ska jeg **ændre nogen ***ting
of course I have to change some things
Com: *((during her talk, C first *closes her eyes, then **gazes up in the air, and ***tilts head slightly. Her talk is accompanied by gestural movements with both hands))*



53 C: *å::h **sæføli ska jeg oss gi mig på ***nogen ting→
and of course I have to compromise on some things
Com: *((*C gazes up, **closes eyes and ***gazes up in the air again; she makes gestures as she talks))*



54 C: **.hh* **men- det ka ik pass jeg ***igen ska gi mig på-(.)****alt**
but it cannot be right that I again have to cave in on everything
Com: *((At her in-breath C's eyes are open, during her talk she first closes her eyes, opens them swiftly and closes them again, and at utterance completion she gazes firmly into her iPhone; she makes gestures as she talks and at turn completion ('alt') *****she first positions her hands on table, ******then places them under the table))*



55 (1.3)



56 C: det ka det ikk
that it cannot be
Com: ((C's gaze is fixed at the iPhone, she shakes her head, her tone of voice is firm, her hands are placed under the table, but at turn completion she lifts right hand up towards face))

57 (0.8)
Com: ((C wipes her eyes))



58 S: *nej*
no

Excerpt 4: *"standing my ground"*.

In excerpt 4, C regains emotional control, which is first and foremost marked by a strong voice quality and the absence of wobbly staccato-like talk. Further, in lines 52-53 she utters a verbal optimistic stance to her own future behaviour, acknowledging that she must make certain compromises in her relationship. This is a standard exit from a troubles talk sequence [48]. In this phase she shifts between closing and opening eyes, gazing up in the air, and her talk is accompanied by repetitive hand gestures (not visible in the transcript). Having made concessions on her own future desirable conduct, she concludes: "but it cannot be right that I again have to cave in on everything" (l. 54). This stance-taking utterance is preceded by a deep in-breath and accompanied by alternating open and closed eyes, repetitive hand gestures, stressed utterance of distinct verbal expressions, and an occasional cut-off. At turn completion, she places her hands first on the table, then under the table (only visible on her side of the screen and not visible in the transcript) and gazes firmly on her iPhone, thereby visibly showing P that she is done. In the long silence in line 55, C keeps eyes firmly on the iPhone, and eventually self-selects as next speaker, upgrading her stance on not 'caving in', this time accompanied by the fixed gaze on the iPhone, head shaking, and hands placed under the table. Combined, her verbal utterance as well as her visual and embodied conduct, reflect her stance on the issue. She stands her ground. One remaining sign that she is still affected is her wiping gestures in the ensuing silence (l. 57). Eventually, the response that C's turn invited for (back in line 54) is delivered by P in line 58: a soft affirmative response, which aligns with C's evaluative stance on standing her ground.

Conclusively, C has managed to come past both the breaking point on her own and to manage to stand her own ground whilst exploring her emotions, without the aid of a comforting touch or any other type of physical or verbal intervention from P. P could have steered the interaction in another direction on several occasions but at no point did she. Rather, she produces continuers at sequentially relevant positions to support the citizen in her troubles talk, and by doing so she supports the citizen in steering through a phase of emotional distress. Had this been a face-to-face encounter, the therapeutic touch could potentially have been drawn upon or called for as an interactional resource by either one to comfort and ease C in handling her emotional distress. However, the video-mediated context allows both participants to employ other interactional resources. Despite what we could possibly assume to be constraints of the video-mediated format, such as the talking heads format [55], lack of mutual eye contact [56], latency issues [57] or haptic issues [58], none of the participants make such constraints relevant in their interaction. On the one hand, the support-contact person orients to C's emotional state as it accumulates by simply being present, and aids C to explore and overcome her emotional distress, by using a few continuers and by using silence as a resource. She is just present. C, on the other hand, explores her emotions and regains control, without the comforting touch.

4. Discussion

Verbal utterances, breathing cues, prosody, face, gaze, and embodiment are interactional resources that each contribute in distinct ways to instances of affect and emotional stance [49, 51, 52, 53, 54, 59]. Concretely, features of crying often manifest themselves in a number of ways: (a) silence, (b) whispering /mouthed talk, (c) sniffs, (d) elevated pitch, (e) wobbly voice, (f) aspiration, (g) sobbing, (h) creaky voice, (i) staccato, (j) swallowing, and (k) throat clearing. Other features of displaying upset may include trembling face and/or hands, touching eyes or face, looking down or hiding one's face, turning away, and flushed appearance, e.g. around eyes and nose [50, 297-299]. Accordingly, identifying features of crying is not straight-forward as the signs often appear in distinct order and /or mark different stages of emotional distress [52].

In our data, evidence shows that verbal and vocal cues (wobbly staccato-like utterances and deep in-breaths), visual cues (in the form of head tilts, gaze shifts, open-closed eyes, mouthy movements, and tear-wiping gestures) together with auditive cues (long verbal silences) are all made relevant by the participant crying in this video-mediated recovery encounter. As such, the mere act of crying resembles and aligns with state-of-the-art studies on crying. The emergence of crying is seen as a stepwise transition, where each cue has a distinct function in the process, not seen as an action in their own right but mostly found to be intertwined with, interfering with, replacing, emphasizing, or adding a dramatizing touch to talk [52, 207]. So, what are the interesting findings?

First of all, we see the sequential act of crying in its entirety and how this is collaboratively accomplished in a video-mediated setting. Next, from the practice in and out of the Danish drop-in centres, we know that touching as a resource to help citizens overcome emotional distress is omnipresent. We also know that both professionals and citizens problematize the shift to a digital context because physical touch is no longer a viable course of action. But how can we know that this video-mediated crying

© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

sequence differs from what we could find in a physical setting? To contrast with the emergent crying sequence analysed above, an example from a physical recovery meeting with *the same citizen* is now being reported on (transcript not included here). In this physical setting, C displays features of emerging crying. The proof procedure lies in the fact that crying in this interaction is discontinued. C's emerging distress is kept in control by a professional through a verbal utterance "Are you ok", followed by a comforting touch on the knee, after which he steers the interaction in another direction, and C stops crying. In our video-mediated encounter presented above, the professional likewise produces a verbal utterance as crying is emerging: "Do you become a bit affected now". But here, the follow-up actions by the professional are made up of verbal continuers and long silences, i.e., simply being present in the situation. She does not steer the interaction elsewhere. And this marks the turning point for allowing the citizen to fully explore and overcome her emotions.

The digital context is arguably partially deterministic for the way the participants collaboratively handle emotion and affect in-situ [49, 60]. Yet, this digital contingency allows for the professional to support a mentally ill citizen to manage her emotional distress without drawing on the therapeutic touch nor other types of verbal comforting actions. The case in point is that the need for comfort is not made relevant by either participant in this video-mediated interaction.

In recent years, a growing body of research on video-mediated interaction has appeared [e.g. 61, 62, 63, 64, 65, 66, 67, 68]. Some notable studies are arguably the talking-head phenomenon [55], (lack of) mutual eye contact and gaze direction [55], and fractured ecologies [60], which amongst other things, it is argued, may result in a fragile shared interactional zone [56]. Further, in a recent review on video-mediated studies, Mlynář et al. [58], point to haptic restrictions and the need for visual awareness and control as drivers for developing future video-communication technologies. This strikes a chord with this study's phenomenon of interest, albeit in a contrasting light.

In a video-mediated recovery meeting, where the very nature of exploring and managing emotional distress becomes the locus, minimal continuers together with subsequent long silences, where the support-contact person is simply digitally present, support the citizen to explore her emotions, and normalize the act of crying [53].

Recovery and rehabilitation work historically stem from a social-psychiatric practice that relies on physical co-presence. The therapeutic touch with its inherent comforting and prosocial features - with which both practitioners and citizens place so much importance - may not always be the recourse in situations where *the very nature* of emotion is the professional locus. From a recovery-oriented perspective, this collaborative accomplishment of 'online' crying is potentially interesting as a therapeutic tool, exactly because of the video-mediated constraints. Mentally ill persons challenged by managing their emotions may benefit from the video-mediated recovery meeting as it prevents the professional from physically comforting the citizen. It facilitates a setting where the citizen can explore and overcome emotional distress. Seen in the light of rehabilitation work, the video-mediated encounter can potentially assist the professional in helping the citizen towards more autonomy (i.e., managing own feelings); developing a new social role (i.e., a competent member of society who can manage one's own affairs in professional digital encounters). And overall, it may assist the citizen towards better life quality.

5. Acknowledgments

I am deeply grateful to the professional(s) and citizen(s) who consented to participate in this research project and granted access to a professional practice which is not easily granted access to. I am especially grateful to the citizen submitting herself to scrutiny in the act of crying. Further, I am indebted to The Velux Foundation and Professor Mie Femø Nielsen, Copenhagen University, for financial support for this research.

6. Transcription notations

Symbol	Description
∞baby∞	Wobbly voice – enclosed by tildes
Hhhelp	Aspiration in speech – an "h" represents aspiration
Huhh . hhih Hhuyuuh >hhuh<	Sobbing – combinations of "hhs", if sharply inhaled or exhaled, they are enclosed in the "greater than/less than" symbols (> <)
(2.5)	Silence – number in parentheses represent silence in tenth of a second)
.h	Hearable inbreath
h.	Hearable outbreath
Word-	Hearable cut-off of word
°Word°	Creaky voice or staccato voice. Two asterics surround the utterance
↑	Upward local intonation
↓	Downward local intonation
↘	Global downward intonation
↗	Global upward intonation
[word]	Top begin-end overlapping talk
[word]	Bottom begin-end overlapping talk
°okay°	Low tone of voice/soft voice – marked by single degree signs
☺word☺	A smile voice
((words)) *((words))	Double bracket signs are the analyst's comments on visible actions/facial expression/embodied conduct of a participant. If asterics are provided, They signify the sequential position of the visible action.
[image] *	An asterics below an image indicates and correlates with an asterics marked in a distinct sequential position in the transcription. If more images are included, additional asterics are included: **; ***, ****, *****, ******
°∞okay∞°	Soft and wobbly voice
Heh- heh-	Nervous laughter – staccato like
ΔwordΔ	Faster talk
=	Latched talk

7. References

1. Herheim, Å. (2008). *Socialpsykiatri –en lærebog.* Academia.
2. Larsen, A. (2002). *Den Socialpsykiatriske Relation—Et forsøg på en bestemmelse af dens karakteristika og betingelser.* Kandidatafhandling. Københavns universitet.
3. Brandt, P. (1996). *Socialpsykiatri - psykiatri på humanistisk grundlag.* Dafolo Forlag.

4. Juliussen, F.B. (2013). *Fra klinik til borgerens livsverden. Paradigmeskift, recovery og rehabilitering.* Presentation, Dansk rehabiliteringskonference, 30/10-13.
5. Karpatschof, L. (1980). *Fra internering til selvorganisering*, Dansk Psykologisk Forlag.
6. Voetmann, C. (2016). Sådan arbejder du med berøring i hverdagen. F*agbladet FOA*, nr. 04, 26-27.
7. Voetmann, C. (2017, 6 Feb). *Charlotte Voetmann: Charlotte Voetmann, neuropædagogisk konsulent, Hjerne & Sundhed* [Video file]. Retrieved 15 April 2020 from https://youtube.com/watch?v=kYKGtzhFaHE&feature=youtu.be
8. Moberg, U. K. (2010). *Afspænding, ro og berøring*. Akademisk forlag.
9. King, A. (2011). Touch as Relational Affirmation. In: K. White, O.B. Epstein, D. Landi (Ed), *Attachment: New Directions in Psychotherapy and Relational Psychoanalysis, Vol. 5*, 108-124.
10. Bickmore, T, W., Fernando, R., Ring, L. & Schulman, D. (2010). Empathic Touch by Relational Agents; in *IEEE Transactions on Affective Computing,* vol. 1, no. 1, 60-71.
11. Lapierre, A. (2003). "From Felt-Sense to Felt-Self: Neuroaffective Touch and the Relational matrix. *Psychologist-Psychoanalyst* Vol. 23 No 4, 43-45.
12. Kauffmann, T. L., & Rosager, H.L. (2019). Sundhedsfaglig praksis: 'varme hænger og kold teknologi". *Klinisk Sygepleje*, Universitetsforlaget, Vol. 33, No 4, 304-315.
13. Garfinkel, H. (1964). "Studies in the routine grounds of everyday activities", *Social Problems*, vol. 11, no. 3, 225–250.
14. Goodwin, M. H. (2017). Haptic Sociality. The Embodied Interactive Constitution of Intimacy Through Touch. In C. Meyer, J. Streeck, & J.S. Jordan (Eds.), Intercorporeality: *Emerging Socialities in Interaction*. Oxford: Oxford University, 73-102.
15. Goffman, E. (1971). *Relations in Public: Microstudies of the Public Order*. New York: Basic Books.
16. Hillewaert, S. (2016). "Tactics and Tactility: A Sensory Semiotics of Handshakes in Coastal Kenya." *American Anthropologist* 118 (1): 49-66.
17. Stuart, S. A. J. (2012). "Enkinesthesia: The Essential Senuous Background for Co-Agency." In *Knowing Without Thinking: Mind, Action, Cognition, and the Phenomenon of the Background*, edited by Zdravko Radman, New York: Palgrave Macmillan, 167-186.
18. Nishizaka, A. (2007). "Hand Touching Hand: Referential Practice at a Japanese Midwife House." *Human Studies* 30 (3), 199-229.
19. Throop, C. J. (2012). "On the Varieties of Empathic Experience: Tactility, Mental Opacity, and Pain in Yap". *Medical Anthropology Quarterly* 26(3), 408-430.
20. Cekaite, A. & Holm K. M. (2017). The Comforting Touch: Tactile Intimacy and Talk in Managing Children's Distress. *Research on Language and Social Interaction*. 50(2), 1-19.
21. Asmuss, B. (2024). The stable organization – routines in organizations: *Book 12- the international business interaction series*. Kindle Edition.
22. Erhvervsministeriet. (2017). *Redegørelse om digital vækst.* Retrieved 17 October 2020 from https://em.dk/media/9710/05-04-redegorelse-om-danmarks-digitale-vaekst-2017.pdf
23. Digitaliseringsstyrelsen. (2015). *Afprøvning af fremtidens velfærdsteknologiske løsninger*. In digst.dk.
24. Arminen, I., Licoppe, C. & Spagnolli, A. (2016). Respecifying mediated interaction. *Research in Language and Social Interaction*, 49, 290-309.
25. Gravengaard, G., & Rimestad, L. (2014). Observationsstudier. In D. N. Hopmann, & M. Skovsgaard (Eds.), *Forskningsmetoder i journalistik og politisk kommunikation*. Hans Reitzels Forlag, 141-163.
26. Goodwin, C. (2010). Multimodality in human interaction. *Calidoscópio*, 8 (2), 85-98.
27. Streeck, J., Goodwin C., & LeBaron, C. (2011). *Embodied Interaction: Language and Body in the Material World*. Cambridge: Cambridge University Press.
28. Stivers, T. & Sidnell, J. (2005). Introduction: Multimodal interaction. *Semiotica, 156 (*1/4).
29. Due, B. L. (2017). *Multimodal interaktionsanalyse. Med videoetnografisk dataindsamling.* Samfundslitteratur.
30. Garfinkel, H. (1967). Studies in Ethnomethodology. Englewood Cliffs, NJ: Prentice-Hall.
31. Garfinkel, H. (2002). *Ethnomethodology's Program: Working Out Durkheim's Aphorism*. Lanham: Rowman & Littlefield.
32. Sacks, H, L., Schegloff, E.A. & Jefferson, G. (1974). A simplest Systematics for the Organisation of Turn-Taking for Conversation. *Language, 50*(4), 696-735.
33. Heritage, J. (2008). 'Conversation analysis and institutional talk: analysing data. In *Qualitative research: theory, method, and practice.* Silverman, D (ed.) London: Sage, 222-245.
34. Nevile, M. (2015). The Embodied Turn in Research on Language and Social Interaction, *Research on Language and Social Interaction,* 48:2, 121-151.
35. Goodwin, C. (2012). The co-operative, transformative organization of human action and knowledge. *Journal of Pragmatics* 46 (1): 8-23.
36. Mondada. L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics*, 65, 137-156.
37. Jefferson, G. (2004). Glossary of transcript symbols with an Introduction. *In:* G.H. Lerner, ed. *Conversation Analysis: Studies from the first generation.* Philadelphia: John Benjamins, 13-23.
38. Hepburn, A. (2010). Crying: Notes on Description, Transcription, and Interaction. *Research on Language and Social Interaction. 37*(3), 251-290.
39. Steensig, J. (2009). *Medie- og kommunikationsteoretisk leksikon.* Kolstrup, S., Agger, G., Jauert, P. & Schrøder, K. (red.). 1. Udgave udg. Frederiksberg: Samfundslitteratur, 518-519.
40. Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality, *Research on Language and Social Interaction,* 51:1, 85-106.
41. Laurier, É. (2014). The graphic transcript: Poaching comic book grammar for inscribing the visual, spatial and temporal aspects of action. *Geography Compass*, *8*(4), 235–248.
42. Laurier, É., & Back, T. B. (2023). Recurrent problems and recent experiments in transcribing video. In *Routledge*

© 2024 Jeanette Landgrebe. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/).

eBooks, 245–264.
43. Liberman, K. (1981). Understanding Aborigines in Australian Courts of Law. *Human Organization*, *40*(3), 247–255.
44. Sommer, M. (2017). *Mental health among youth in Denmark*. Retrieved 14 February 2020 from http://nordicwelfare.org/wp-content/uploads/2017/10/danmark_webb-1.pdf.
45. The Ministry of Health. (2017). Healthcare in Denmark. An overview. Retrieved 27 May 2020 from https://www.sum.dk/English/~/media/Filer%20-%20Publikationer_i_pdf/2016/Healthcare-in-dk-16-dec/Healthcare-english-V16-dec.ashx
46. Landsforeningen af væresteder. (2005). *Samfundets værested. Fra omsorg og lavtærskel til samvær og aktivitet.* Retrieved 15 May 2020 from https://vaeresteder.dk/media/1208/samfundets-vaerested.pdf.
47. Thorsager, L., Børjesson, E., Christensen, I., Pihl, V. (2007). *Metoder i socialt arbejde. Begreber og problematikker.* Socialforskningsinstituttet. København 2007.
48. Jefferson, G. (1984). On stepwise transition from talk about a trouble to inappropriately next-postitioned matters. In J.M. Atkinson and J. C. Heritage (Eds). *Structrues of social actions: Studies of conversation analysis.* Cambridge University Press, 191-222.
49. Peräkylä, A. & Sorjonen, M-L. (2012). *Emotion in Interaction*. Oxford Studies in Sociolinguistics.
50. Hepburn, A. (2004). Crying: Notes on description, transcription, and interaction. *Research on Language and Social Interaction, 37*(3), 251–290.
51. Laakso, M. (2014). Aphasia Sufferers' displays of Affect in Conversation, *Research on Language and Social Interaction*, 47: 4, 404-425.
52. Hepburn, A. & Potter, J. (2012). Crying and Crying Responses. In: Peräkylä And Sorjonen (Eds); *Emotion In Interaction*. Oxford University Press, 195-211.
53. Reber, E. (2012). Affectivity in interaction. In *Pragmatics and beyond*. John Benjamins Publishing, 3–8.
54. Ogden, R. (2021). "Swallowing in Conversation." *Frontiers in Communication*, vol. 6, pp XX
55. Licoppe, C. & Morel, J. (2012). Video-in-interaction: "Talking Heads' and the multimodal organization of mobile and skype video calls. *Research on Language and Social Interatction*, 45, 399-429.
56. Fornel, M. de. (1996). The interactional frame of videophonec exchange. Rëseaux: Fr. J. Commun., 4, 47-72.
57. Seuren, M.L., Wherton, J., Greenhalgh, T., Shaw, E.S. (2021). Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics* 172, 63-78.
58. Mlynář, J. González-Martínez, E. & Lalanne, D. (2018). Situated Organization of Video-Mediated Interaction: A Review of Ethnomethodological and Conversation Analytic Studies. *Interacting with Computers,* 1-13.
59. Goodwin, M. H., Cecaite, A., & Goodwin, C. (2012). Emotion as stance. In *Oxford University Press eBooks*, 16–40.
60. Luff, P., Heath, C., Kuzuoka, H., Hindmarsh, J. Yamazaki, K. & Oyama, S. (2003). Fractured ecologies: creating environments for collaboration. Human- *Computer Interaction*, 18, 51-84.
61. Arminen, I., Licoppe, C. & Spagnolli, A. (2016). Respecifying mediated interaction. *Research in Language and Social Interaction*, 49, 290-309.
62. Vasilyeva, Z. (2013). Video-mediated communicative interaction: an analysis. *Forum Anthropol. Culture*, 117-148.
63. Mondada, L. (2015). Ouverture et prëouverture des rëunions visophoniques. *Réseaux,* 6, 39-84.
64. Licoppe, C. (2017). Showing objects in Skype video-mediated conversations: from showing gestures to showing sequences. *Journal of Pragmatics*, 111, 63-82.
65. Rintel, S. (2013). Video calling in long-distance relationships: The opportunistic use of audio/video distortions as a relational resource. *The Electronic Journal of Communication.*
66. Hjulstad, J. (2016). Practices of organizing built space in videoconference-mediated interactions. *Research on Language and Social Interaction,* 49, 325-341.
67. Due, B.L., Lange, S.B., Nielsen, M.F. Toft, T.L.W., Landgrebe, J., Nielsen, A. M.R., Hassert, L.O. (2019). Den andens ansigt: Afdækning af deltagerorientering via multimodal interaktionsanalyse. *Nys,* 56, 10-51.
68. Stommel, J.P.W., Goor, H., Stommel, W.J.M. (2020). The Impact of Video-Mediated Commmunication on Closed Wound Assessments in Postoperative Consultations: Conversation analytical Study. *Journal of Medical Internet Research*. Vo. 22, iss.5.