IJVS - INTERNATIONAL JOURNAL OF VOICE SCIENCES

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

From Voice Signals to Voice Maps

Von Stimmsignalen zu Voice Maps

Sten Ternström[1] and Peter Pabon[2]

[1]*Division of Speech, Music and Hearing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden*
[2]*Voice Quality Systems, Utrecht, The Netherlands (retired)*

Received 1 September 2025, accepted 30 October 2025

Abstract

This article is intended as an introductory tutorial for technically inclined clinicians, vocologists and voice pedagogues who want to understand the principles and potentials of voice mapping. Voice mapping has its origins in the Voice Range Profile, or phonetogram, but it is less concerned with the extremes of the voice range, and more with what happens *within* a relevant range of the voice. It is a voice instrumentation paradigm that is intended to improve the evidential value of voice measurements. It exposes and automatically accounts for the strong co-variation that most voice metrics exhibit with fundamental frequency and sound level. Very many data points are automatically collected in a short time, and their means are mapped by colour onto maps. This results in a robust representation of voice status and function. While individual voices are very different, a voice map's appearance is reproducible within individuals. Comparing maps across interventions gives rich information, even on subtle changes in a voice. Moreover, by statistically clustering multiple metrics, phonation types can be identified and mapped automatically, thereby enhancing clinical relevance, and facilitating a deeper understanding of voice data.

Abstrakt

Dieser Artikel ist als Tutorial für technisch Interessierte gedacht, die die Prinzipien und Möglichkeiten des Voice Mapping verstehen möchten. Voice Mapping hat seinen Ursprung im Voice Range Profile oder Phonetogramm, befasst sich jedoch weniger mit den Extremen des Stimmumfangs als vielmehr mit dem, was innerhalb eines relevanten Stimmbereichs geschieht. Es handelt sich um ein Paradigma, das Stimmmessungen verbessern soll. Es deckt die starke Kovariation auf, die die meisten Messmethoden mit den Parametern Grundfrequenz und Schallpegel aufweisen. Dies führt zu einer robusten Darstellung des Stimmzustands und der Stimmfunktion. Während sich einzelne Stimmen stark unterscheiden, ist das Erscheinungsbild einer Voice Map innerhalb einer Person reproduzierbar. Darüber hinaus können durch statistisches Clustering mehrerer Metriken Phonationstypen automatisch identifiziert und abgebildet werden, wodurch die klinische Relevanz erhöht und ein tieferes Verständnis von Daten ermöglicht wird.

Keywords:

Voice map, voice range profile, voice measurement, electroglottography, clinical evidence

Schlüsselwörter:

Voice Map, Stimmfeldmessung, Stimmparameter, Elektroglottographie, klinische Evidenz

INTRODUCTION

As clinicians or scientists, we know that the human voice is a complex system, and we try to find ways of quantifying its functioning and status. Describing such a system necessarily requires a lot of information. Inconveniently, our human vision and cognition are terrible at reading and understanding long tables of numbers – but they are remarkably good at quickly assessing scenes and images, such as maps.

In speech and song, the acoustic properties of vocal sounds change continuously, thereby encoding the message. This means that measuring these properties at any single instant will give only an obscured view of voice function, a view that is largely determined by the task performed by the participant. The considerable variability associated with the task type and how participants respond has been challenging to voice science. A legacy approach to deal with this variability has been to eliminate as much of the variability as possible, by asking the participant instead to produce sustained vowels at a fixed loudness and pitch. But in so doing, one also discards a great deal of information, information that clinicians,

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

performers and pedagogues readily perceive and use to build a mental map of what a given voice can and cannot do. So can computers do this, too? Yes – voice mapping is an instrumentation paradigm for acquiring voice signals and then collating, condensing and visualising information from the signals onto images or maps. If the acquisition is observed in real time, the clinician also sees how well the participant is able to *control* their voice. The result is a quantified, summarising representation of the participant's voice; one that is richer and more interpretable than conventional representations.

The aim of this article is to explain what voice maps are, and what they can show us by following the full path from voice signals to voice maps. To set the scene, we briefly revisit the underlying concepts of signals, metrics, distributions and time scales. Then, we consider how voice maps are made, how mapping different entities reveals co-variation, and how that co-variation leads us to use statistical clustering as a tool for clinical voice classification. Along the way, we will show examples of mapping the effects of some voice interventions. To stay in the flow, some details and examples, that are not strictly necessary for the general ideas, are deferred to boxes and tables.

1 THE VOICE AS A MACHINE

The voice organ can be thought of as a machine. Like any other machine, it is designed to run in a certain range of operating conditions, and it will work best when it is free to run within that range. As an analogy, consider a car engine: it is designed to run at a range of speeds, and if it cannot, that indicates a problem. Even if there is nothing wrong with it, a poorly controlled engine may find itself running in some inefficient part of its range. And even within its intended range, strange behaviour from the engine can indicate that the normal relationship between the controls and the desired motion is disrupted, or that parts of the machine are out of order. For the voice "mechanic", or clinician, modern voice mapping has the threefold function of (a) determining the range of a voice, (b) documenting which parts of that range that the user currently tends to occupy, and (c) showing aspects of the operation of the underlying mechanism, in that context.

Signals

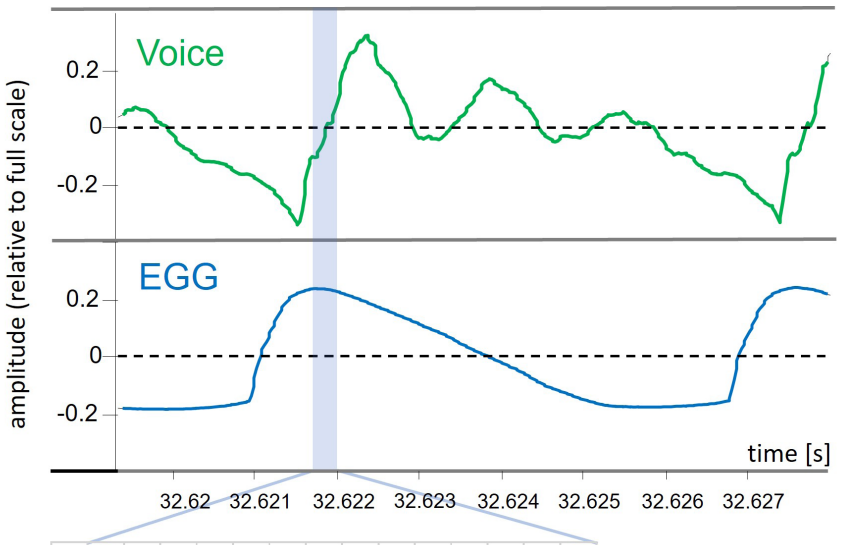
When making measurements of a voice, there are several possible physical entities to choose from. The most common one is of course the sound that we hear and can record using a microphone. A microphone senses variations in the air pressure, and converts them into a changing electrical voltage that is proportional to the pressure at each instant in time. We say that the microphone *transduces* the pressure into an electrical analogue of the pressure. Several more signals and their origins are listed in Table 1.

For analysis by computer, the electrical signals from microphones and other transducers are first digit-

Signal	Method name	What the signal represents
Sound	Audio recording	The instantaneous sound pressure, which is a tiny deviation from the atmospheric pressure. Running averages of its energy, repetition frequency, regularity, and spectral distribution, are perceived by our sense of hearing as loudness, pitch, pitch salience, and timbre, respectively.
FPG	Flow pneumotachography	The instantaneous rate of airflow past the lips, from which the airflow past the glottis can be estimated by inverse filtering. Measuring the airflow in the glottis *in vivo* directly is very difficult.
EGG	Electroglottography	The instantaneous trans-laryngeal electrical conductance (which is the inverse of the electrical resistance). From the resulting current, variations in the relative medial contact area of the vocal folds can be inferred. [1][2]
PGG, NIR PGG	Photoglottography	The amount of visible or near-infrared [3] light passing through the open glottis from a light source to a photo-sensor, from whose signal the variations in the area of the glottal opening can be inferred. These variations can be represented as the glottal area waveform, GAW.
ACC	Accelerometry	The acceleration (vibration) of the skin, usually from a transducer placed on the neck just below the thyroid prominence. This signal can be processed so as to estimate several phonatory metrics. [4]
HSV	High-speed video-endoscopic imaging	A very rapid sequence of images of the larynx, taken from above. This gives a torrent of data that can be interpreted in many ways. Commonly, the GAW is computed by image processing.
EMG	Electromyography	Electrical nerve impulses that control muscle activations.

Table 1. Some types of signals that are relevant to phonation and can be collected from a voice. All except the first two have the advantage of picking up phonatory signals before they are subjected to articulation by the vocal tract, which adds another tier of information and variability. The last two are rather invasive.

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20



Time (s)	EGG	Voice
32.62171	0.235372	-0.10234
32.62174	0.235097	-0.08558
32.62176	0.234792	-0.07261
32.62178	0.234823	-0.05658
32.6218	0.234731	-0.03063
32.62183	0.233876	-0.00316
32.62185	0.233144	0.015156
32.62187	0.232259	0.013599
32.62189	0.230061	0.019705
32.62192	0.228626	0.044156
32.62194	0.227558	0.066288
32.62196	0.225512	0.089182
32.62198	0.223711	0.119708
32.62201	0.221513	0.15634

Figure 1. Analogue signals are represented digitally as sequences of numbers. The figure shows a very short excerpt of about 1½ cycles of phonation, for about 8 ms in time. Here, the sampling rate is a standard 44100 samples per second. In the table, the two sequences of sample values with a duration of only 0.3 ms corresponding to the shaded part are shown. The time points are implicitly given by the sampling rate, and are not usually stored.

ised into rapid streams of numbers, or *samples*, by an electrical circuit called the analogue-to-digital converter (ADC). This is done at rates of tens of thousands of samples per second for phonatory signals (Figure 1), or hundreds per second for slower physiological signals. The resulting numbers are then stored by the computer, in signal files. This is what a data acquisition system does. Signals that are represented as sequences of numbers are called digital signals.

Metrics

By a procedure that is the reverse of data acquisition, via a digital-to-analogue converter (DAC), we can play back digital signals, and listen to them for subjective assessment. The different recognisable voice qualities that we may wish to represent on maps are all linked to specific signal properties, though sometimes indirectly. To quantify those signal properties, we devise different *metrics* and compute them from the sequences of sample values. Here, we will use the term 'metric' to mean some property or feature of the signal that can be represented by a single computable number, or 'scalar.' For the voice clinic, our hope is that the value of a metric will somehow be relatable to the voice status of the participant. Furthermore, if the values of a metric are used to control some mathematical model that reconstructs the signal, the metric becomes equivalent to a *parameter* of that model. Some examples of metrics (Table 2) are the peak-to-peak amplitude of a flow glottogram (Û), the contact quotient (CQ) of an electroglottogram, and the sound pressure level of an acoustic signal (SPL). Note that the metrics in Table 2 are all physical, not perceptual. For instance, pitch is the main perceptual correlate of the physical entity of f_o. In psychoacoustic research, metrics have been defined also for pitch and many other perceptual quantities, but we will not be concerned with those here.

Distributions of signals

It is interesting to note that the information in a voice signal lies not in the individual sample values

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

Symbol	Unit	Name	Computed as	Basis
ACOUSTIC SIGNAL				
SPL or L_P	dB	sound pressure level	$L_p = 20 \times log_{10}\left(\frac{\text{RMS of signal}}{\text{ref. pressure}}\right)$	energy
f_o	Hz or semitones	fundamental frequency	cycles per second, or *log* (ratio to a reference)	period
CF	- , or dB	crest factor	ratio of peak amplitude to RMS amplitude	energy
{ » }	%	jitter	» variability in period times	entropy of the period
{ » }	dB or %	shimmer	» variability in period amplitudes	entropy of the energy
SB, α, { SPR }	dB	spectrum balance	» ratio of (high-frequency energy) to (low-frequency energy)	*energy*
{ H1-H2 }	dB	"H1-H2"	the ratio of powers between harmonics 1 and 2 in the acoustic spectrum	*energy*
HNR	dB	harmonics-to-noise ratio	ratio of harmonic energy to noise energy	*entropy*
CPP	dB	cepstral peak prominence	(approximately) the periodicity of the log power spectrum	*entropy*
EGG SIGNAL				
CQ_{EGG}	-	contact quotient	fraction of cycle for which vocal folds are in contact, based on » segmentation	glottal adduction
Q_{ci}	-	contact quotient by integration	the area below the normalised EGG pulse	glottal adduction
Q_{Δ}	-	normalised peak dEGG	the maximum derivative[i] of the normalised EGG pulse	VF contacting
{ CSE }	-	cycle-rate sample entropy	cycle-to-cycle variability of the normalised EGG pulse [5]	entropy
GLOTTAL FLOW SIGNAL				
CQ { OQ }	-	closed/open quotient OQ = 1 – CQ.	fraction of cycle for which the glottis is open/closed, based on » segmentation	wave shape
{ NAQ }	-	normalised ampl. quotient	ratio of AC amplitude to the negative peak flow	energy
{H1*-H2*}	dB	"H1-H2"	the level difference between harmonics 1 and 2 in the source spectrum	*energy*
HRF	dB	harmonic richness factor	the ratio of power in harmonics 2-10 to the power in harmonic 1	*energy*

Table 2. Some examples of voice metrics and how they are derived from signals. Most metrics tend to increase with increasing vocal effort, which is the expected direction. In this table, those that conversely tend to decrease are shown with symbols in { braces }. A chevron » means that several definitions occur in the literature. Entities that are dimensionless are shown with the unit '-'. Where Basis is given in italics, the metric is defined in the frequency domain, otherwise in the time domain.

themselves, but in the evolving shapes of the continuous waveforms from which those sample values were taken. Most voice metrics are instead related to one of three very basic signal properties: (1) *energy* or how strong it is, (2) *entropy* or how irregular and unpredictable it is – both can be computed for all signals; and (3) the repetition rate, or *fundamental frequency* (f_o), of the voiced speech segments only. These three must all be computed by first collecting a stretch of signal of some suitable duration, then forming a *distribution* of the samples contained in that interval, and finally computing some descriptive attribute of that distribution. Examples of metrics of *energy* are:

- sound pressure, proportional to the standard deviation[1] of the distribution of the microphone signal;
- acoustic power, proportional to the variance of the distribution of the microphone signal; and
- DC glottal airflow, proportional to the mean of the distribution of the flow glottogram.

In all three cases,[2] the resulting value is a single number, a metric, which summarises aspects of what the signal was like during the chosen time interval.

Time scales

If we are looking at the oscillatory pattern of the vocal folds, then we need a resolution in time that is much finer than the period time of the glottal oscillation, with at least thousands of observations or samples per second (Figure 1). If instead we are

1 Also known as the RMS, or "root of the mean of the square" of the signal

2 The physicist may point out that these three are different expressions of energy: potential energy, energy per unit of time, and kinetic energy.

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

interested in, say, the maximum speed of glottal flow cessation, then that metric is defined only once per phonatory cycle. This reduces the amount of data, from several hundreds of sample points per cycle, to just one number per cycle (Box 2). Indeed, for most metrics of voice signals, the phonatory cycle time is generally the shortest meaningful interval [6]. Its duration ranges from about 15 ms down to 1 ms depending on the f_o.

An example: if we want to compute the SPL cycle-by-cycle, we (or rather, the computer) must find the beginning and end of each phonated cycle in the acoustic signal, then compute the standard deviation (SD) of the digitised samples in the cycle; and express the ratio of this SD to that of a signal with a reference amplitude. If we want the SPL of a sustained vowel (seconds), or of a full afternoon of teaching (hours), we can do the same for the much longer time interval; and perhaps first removing any silent or unvoiced parts. In fact, the time scale of our measurements is very important for what they mean. Different aspects of a voice are associated with different time scales (Table 3).

2 CYCLE-BY-CYCLE ANALYSIS

Vocal fold vibration is a cyclical, repeating phenomenon. It is often desirable to set the analysis time interval to the exact length of each examined phonatory cycle, if the duration can be accurately determined. Such analysis is called 'cycle-synchronous' or (less aptly) 'pitch-synchronous'. A metric can be devised to represent some reduced aspect of the *shape* of one cycle of the periodic signal, such as its maximum slope, or its degree of asymmetry.

Specific aspects of a waveform shape can usually be obtained directly from the distribution of samples across a cycle. For instance, the peak-to-peak amplitude is simply the difference between the maximum and the minimum sample values in that distribution. Or, computing the differences between consecutive sample points of a flow signal yields a series of numbers that are a good approximation of the *time derivative*[i] of the flow. From the distribution of the derivative over the inspected cycle, we immediately find the maximum flow declination rate (MFDR) as the minimum value (because it is negative). The MFDR is a very important metric that is a predictor for several others, including the acoustic output power of the voice.

3 THE FREQUENCY DOMAIN

So far, we have considered signals only in the time domain. By mathematically transforming the signal to the frequency domain, typically using the Fourier Transform, we obtain the *spectrum* of a signal over some suitable time interval. The spectrum represents the distribution of the signal's energy over different *frequencies*, during that time interval. This distribution is closely related to voice timbre. Again, a distribution is not one single number but many, so the spectrum too, is often characterised using scalar metrics taken from its descriptive statistics, such as the spectrum slope, or the spectrum centroid; some more are shown in Table 2. Indeed, very many such spectral metrics have been proposed (Table 2, [8]).

Meta-signals

In practice, we usually compute our metrics from distributions over short time intervals. Now, if we arrange many sequential observations of a metric into a series on a time line, then this series of values let us see how the metric unfolds over time. An example would be a time contour of f_o, perhaps as a representation of speech prosody, or of voice tremor.

Time scale (s)	Voice-related sources of variation	Measurements
10^{-4}	< signal >	< sampling rate >
10^{-3} (ms)	Vocal fold motion	Oscillograms, HSV, phase plots, perturbations, EGG, GAW
10^{-2}	Muscle actions, neural control	EMG, periodicity, f_o, inverse filtering, jitter, shimmer
10^{-1}	Phonation, articulation	Sound pressure level, formants, spectrum sections, tremor, flutter
10^{0} (s)	Effort, prosody	Spectrograms, speech contours
10^{1}	Respiration	Sustained vowels, maximum phonation time
10^{2} (minutes)	Tension/posture, physiological base	Voice range profiles, long-time average spectra (LTAS)
10^{3} – 10^{4} (hours)	Vocal fatigue	Voice dosimetry
10^{6} – 10^{7} (weeks)	Progress of disease, therapy or training	Repeated assessments
10^{9} (decades)	Age, health status	Longitudinal studies

Table 3. Examples of sources of variation in measurements of voice signals and meta-signals, arranged by their time scales in seconds. Adapted from [7].

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

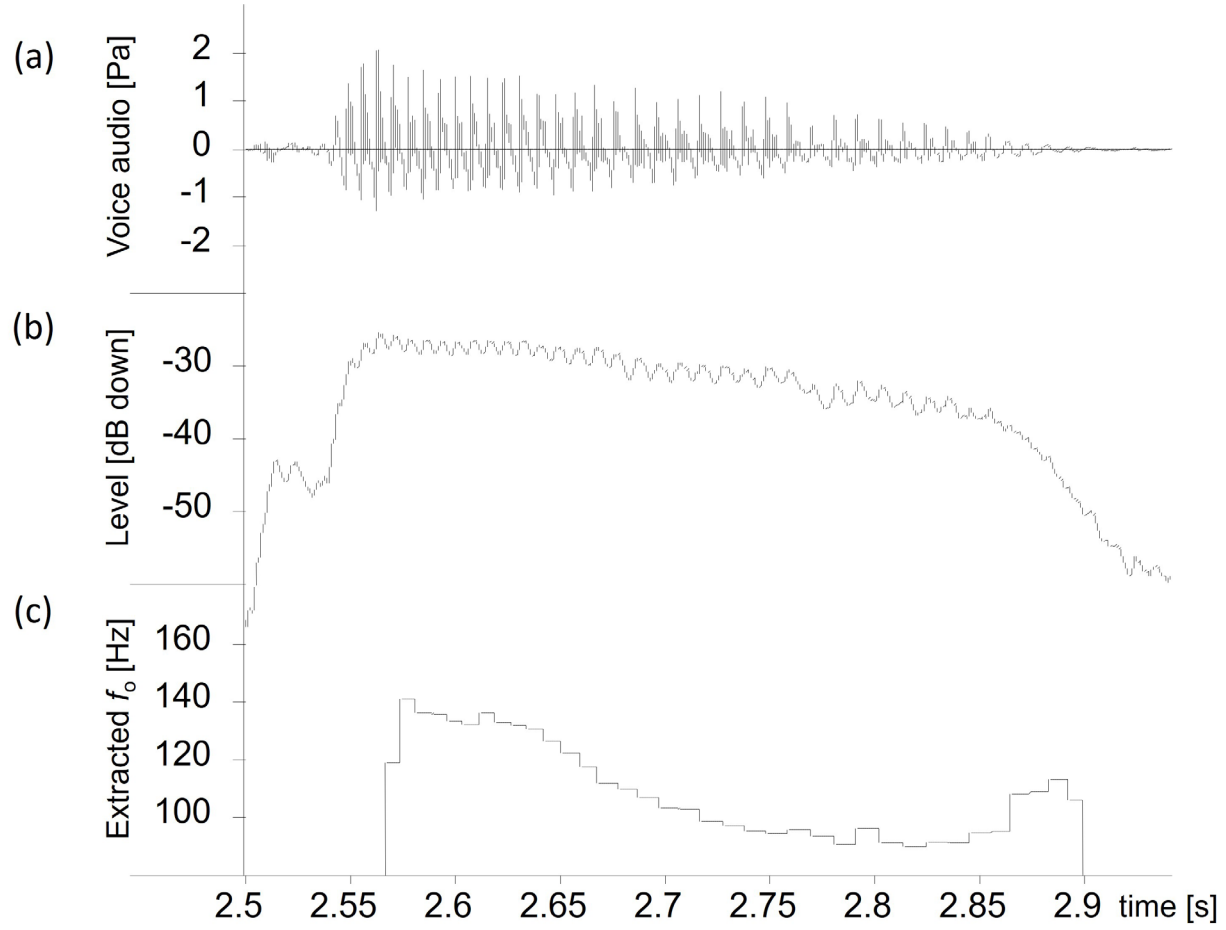


Figure 2. Signals and meta-signals. (a) Voice signal of the Swedish word /tal/ ("number" or "speech"), with clearly visible glottal periods. (b) The signal level of that utterance, estimated by squaring the signal in (a), low-pass filtering it at 50 Hz, similar to a short-term running average, and taking the logarithm of the result. The result can be treated as a signal in its own right, a 'meta-signal'. (c) The fundamental frequency f_o, estimated as a cycle-synchronous measurement of (a). In (c) each step corresponds to the inverted period time of the most recent glottal period. This is an example of a meta-signal that is defined only for voiced parts of the original signal.

Such a series of points can then be treated as new signal, which we might call a 'meta-signal.' The same analyses can be applied to meta-signals as to signals; we just need to make sure that we understand what the results mean. Thus, we compute new distributions of the meta-signals over longer periods of time, to obtain for instance the mean f_o over a reading of an entire text or the severity of a tremor.

Distributions of metrics

Distributions are usually visualised using histograms. If we arrange observations of a metric on its own axis rather than on a time axis, we can make a histogram showing the occurrences of different values of the metric, usually on a relative scale in percentages. If the plotted range extends to all observed values, then the area under the histogram sums to 100%. For instance, we can plot the distribution of the sound level of a voice, in order to see how much of the total time that the speaker has spent in soft or in loud voice.

Because the voice is so variable, the distributions of metrics are usually more relevant to voice assessment than are the metric values at any one instant. We are more interested in the shape of the distributions and their descriptors (mean, variance, skewness etc.) over some extended time than in the value of a metric during one sustained vowel. Consider the three f_o histograms in Figure 3 of normal, raised and nearly shouting speech. The means clearly differ, but the shape of each distribution also tells us something where this voice was phonating, in relation to its operating limits – which are personal, specific to the individual.

Mapping metrics

If many observations of a metric (a dependent variable) are placed onto the axis of another metric

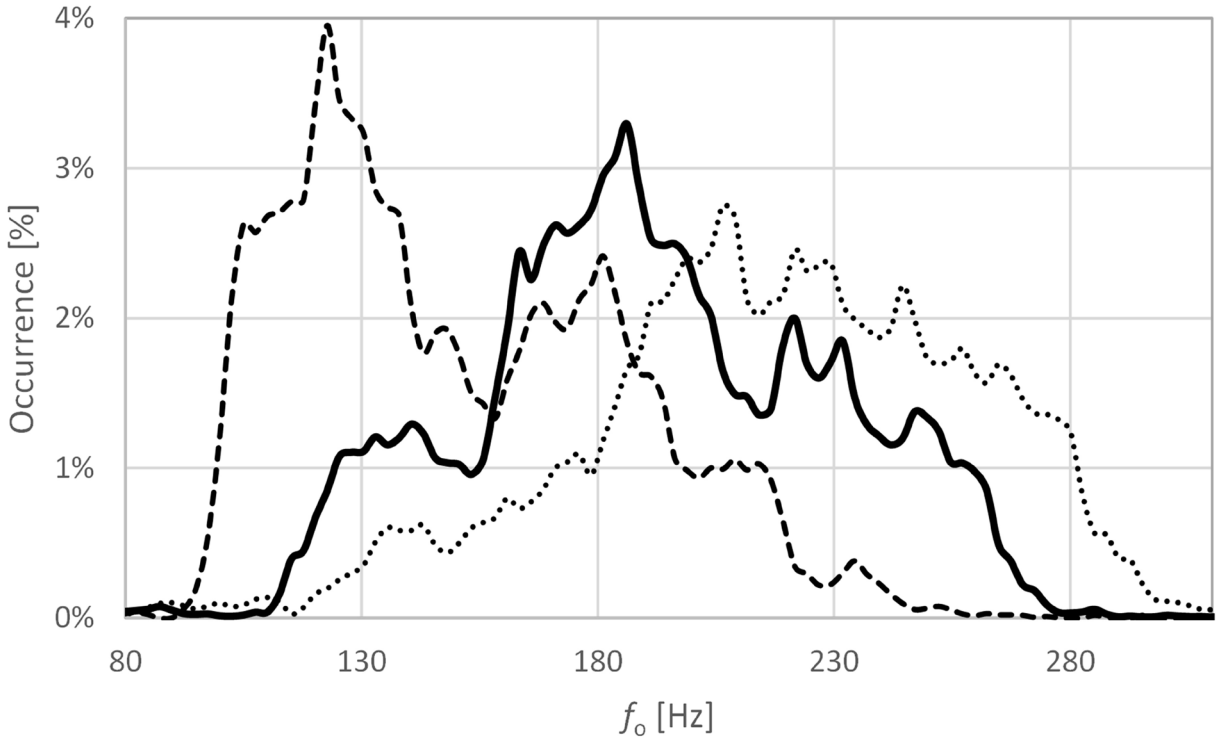


Figure 3. Fundamental frequency (f_o) histograms of a male speaker instructing a group, in quiet (dashed), moderate background noise (solid) and very strong background noise (dotted). The distribution shapes give some information on how the voice is being used (see text). The data are from the study reported in [9].

(an independent variable), we get a cloud of points. If this cloud follows a discernible trend, such that we can draw some descriptive curve through it, then that curve suggests how the first metric varies in relation to the second. For instance, we can see if a metric varies systematically, as the metric 'sound pressure level' (SPL) changes. In Figure 4 we see how a metric called 'CPP' varied when an untrained male speaker vocalised many times from very soft to very loud, at all his possible pitches.

In the middle of the range, where habitual speech resides, the CPP increased by about 5 dB for every 10 dB in SPL, while at the soft and loud extremes, the CPP stayed constant when SPL varied. Given that 40 to 110 dB is a huge range in terms of acoustic power (one to $10^{\left(\frac{110-40}{10}\right)} = 10^7$, or one to ten million), we may fully expect that the voice will behave differently at such very different output levels (Box 5: Logarithmic scales). Importantly, if the curve is clear enough, this hints at the existence of some underlying *function* of voice production that might be captured by a model. Here the term 'function' is applicable in both its clinical and mathematical senses (Box 6: Functions and fields).

4 FRAME-BY-FRAME ANALYSIS

Since the phonatory cycle time is initially unknown, or even non-existent in unvoiced sounds, we usually first examine short fixed-length sections of the signal, or *frames*, to see if we can detect periodicity. For instance, to compute the spectrum of a signal, we choose a time interval or frame, of perhaps 2^{10} = 1024 sample points on a time axis, apply a transform to those data, and obtain a new frame with as many points on a frequency axis instead. Consequently, frame-based metrics are obtained at a regular pace of only once per frame, and not for every glottal cycle. There is usually a trade-off between resolution in time and in frequency; one cannot simultaneously have high resolution in both.

5 LOGARITHMIC SCALES

The semitone scale for fundamental frequency and the decibel scale for sound level are both

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

logarithmic, that is, they express values as ratios. The evolution of our auditory system, and of most other modes of perception as well, has found that ratios are an effective way of covering very wide ranges of variation, such as those from soft to loud sounds and from low to high tones. Nature knows no units, such as meters or Hertz; instead, living creatures obtain the information they need from relative assessments, such as "I am a lot bigger than you are".

In voice mapping, we strive to put *all* metrics on logarithmic scales, that is, we prefer to represent a measured value using its ratio to some agreed reference. This not only enables us to cover the voice range in a single sensible graph, but also affords a very useful invariance with scale. Logarithmic scales offer important advantages for interpreting relations between metrics (see [17], chapters 4 and 5) and they also have consequences for how we think about statistical distributions. For instance, the meanings of variance and standard deviation on a logarithmic scale are different from those on a linear scale.

6 FUNCTIONS AND FIELDS

To mathematicians, a 'function' means a relation that maps the value of an input variable x to an output value y, written as $y = f(x)$, where the function itself is called f. If the output variable y reports a single value, or 'scalar', we can use the function to generate a 'scalar field' (another term for our curve) by observing many values of x and plotting the corresponding y values on a screen or paper. If another function g has *two* input values x and y, but still one scalar output value z, then we write $z = g(x, y)$. The scalar field can then be constructed by making observations at multiple places on x and y, and connecting the resulting points above the plane of x and y, in effect forming a curving surface in 3D. This is how surveyors of landscapes make topographical maps. To visualise the height of the land on a flat paper or screen, the dependent variable z is mapped onto some suitable colour scale. This is also how voice maps work.

To clinicians, 'function' means how the voice organ responds to its owner's intentions, and this too can be seen as a relation between inputs (coming from the brain) to outputs (muscle activations, and ultimately speech or song).

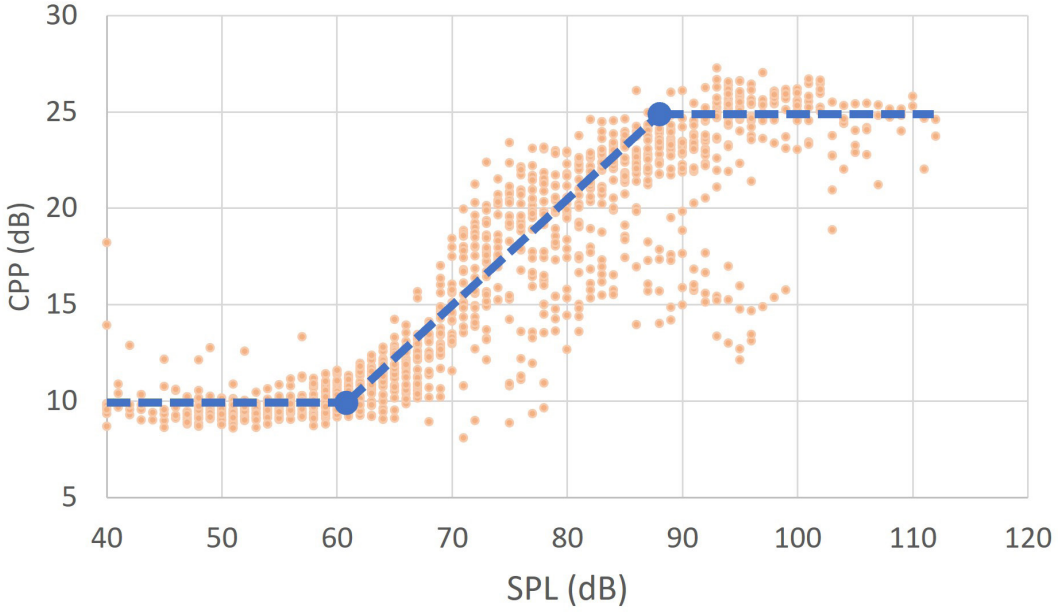


Figure 4. Plotting a metric against another can reveal some underlying function. A metric called 'CPP' varied with voice SPL in an untrained male speaker. CPP stayed constant at about +10 dB in soft voice up to 60 dB SPL, then increased to about 90 dB SPL, and then remained at +25 dB above 90 dB SPL. If we can explain such trends, then something about this voice has been learned. When both axes are in logarithmic units such as dB, linear fits such as the dashed lines shown here are of particular significance (see Box 5: Logarithmic scales).

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

If there is a lot of scatter, and no clear curve can be inferred, we still do not need to be despondent. There could be other systematic influences that also need to be accounted for. If so, we might need a curving *surface* instead of just a curve. This requires *two* independent variables on *two* axes, such as horizontal and vertical, that define a plane. You can see where this is going: if we put f_o on the horizontal axis and SPL on the vertical axis, we get a coordinate system on a plane, which we will call the *voice field*.

The Voice Range Profile

The voice field is the coordinate system of a class of graphs that has variously been called 'phonetograms', 'voice range profiles' (VRP) or, historically, 'Stimmfeld' in German. Before continuing to voice maps, let us first look at a VRP (Figure 5). An untrained male speaker phonated over his full voice range on the vowel /a/, which took about six minutes with appropriate guidance. The VRP shows the accessible range of the participant's voice, that is, the frequencies and sound levels at which he was able to phonate at all. Even flat graphs of this type can be clinically useful, and much has been written about the VRP; for an overview, see [10]. Nevertheless, phonating at the extremes of one's voice is a very unusual and demanding task, which for most voice users requires careful guidance by a therapist.

The Voice Map

Often we are interested in how the voice is working in some smaller, more habitual part of its range. We also wish to see how some metric of interest varies over the chosen range. This principle was first suggested by Pabon and Plomp in 1988 [11], and it forms the foundation of voice mapping. To visualise a metric's values on a flat screen or paper, we typically use a colour scale. The map remains a single graph, but it contains more information than a curve. It shows the dependence of a metric on *two* factors rather than just one. Let us arbitrarily choose the EGG metric Q_Δ [12], and see how its changing values can be represented.[ii] Figure 6 shows four ways of graphing the metric Q_Δ , from the same recording as for Figure 5.[iii] The histogram in Figure 6a shows that Q_Δ ranged from 1 to 18, with the most common value at about 1, with a secondary mode at about 6, and with few values above 10. However, this graph offers no connection to the voice range.

In Figure 6b we see how observations of Q_Δ are distributed over SPL, just as we did for the CPP in Figure 4 (and from the same recording!). The upward linear regression line suggests a weak positive correlation, but that line seems not to be very descriptive of how the points are distributed. In Figure 6c, which instead shows the distribution of Q_Δ over f_o, the regression line now slopes downward, and again it is not very descriptive. It is only from Figure 6d, the voice map of this metric, that we begin to understand

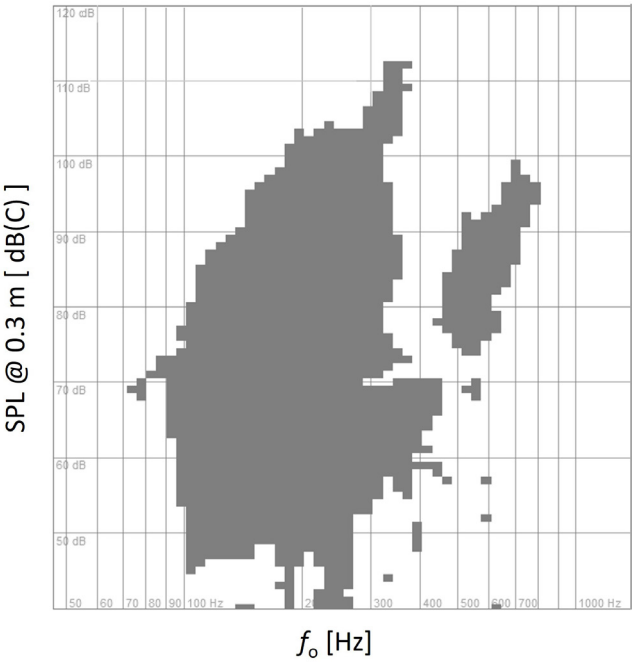


Figure 5. A voice range profile (VRP) of an untrained male speaker. A VRP shows the full range of the participant's voice, that is, the frequencies and sound levels at which he can be elicited to phonate at all. Data adapted from the study reported in [27].

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

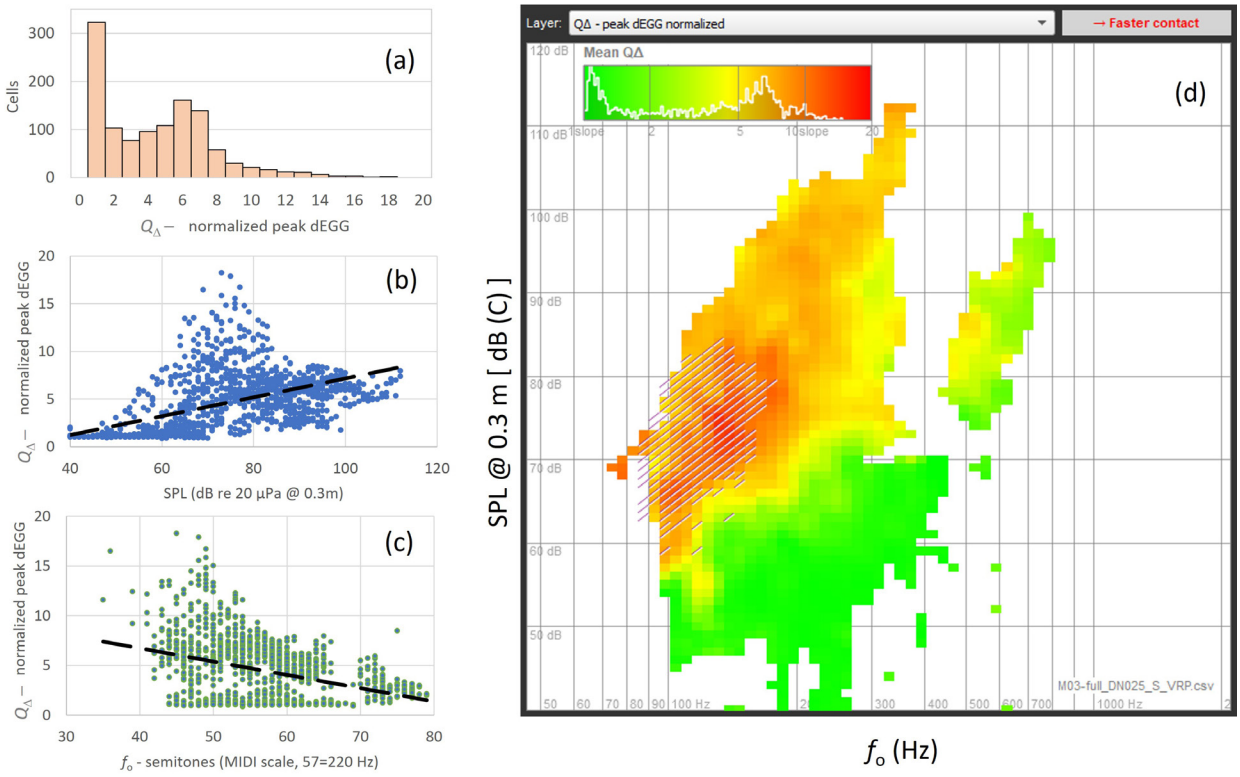


Figure 6. Four ways of graphing a scalar metric, here Q_Δ , from the full voice range of an untrained male speaker vocalizing on the vowel /a/. (a) The histogram of Q_Δ on its own axis, (b) how Q_Δ varied with SPL, with a dashed regression line, and (c) how Q_Δ varied with f_o. All three are combined in (d), which shows how Q_Δ varied over the voice field. The cell colour shows the mean Q_Δ for all visits to a cell. The hatched area shows the range of this speaker reading aloud the "Rainbow Passage" three times. The colour bar contains a histogram similar to that in (a) but on a log scale. A threshold of at least 6 cycles per cell was imposed for all these graphs.

what is going on. The map is colour-coded, and each occupied bin or *cell* acts like a pixel in an image. Here a green cell means that the average value of Q_Δ in that cell was at its minimum value of 1, while an orange cell means that the average Q_Δ was 10.

Unlike the scatter evident in (b) and (c), the voice map reveals an essentially smooth surface, as colour transitions (d) between adjacent cells are progressive. So, the metric typically changes *gradually* with f_o and SPL, and the shape of the surface is telling us something about how this voice works. We discern a steepest gradient in the transition from green to yellow that runs diagonally through the main 'island'. In maps of Q_Δ , this feature marks the onset of vocal fold contacting. A separate 'island' to the 'east' represents the speaker's limited falsetto range, as confirmed by inspection of additional metrics (or by listening!).

We can now draw a very significant conclusion: in Figure 6, *the vertical scatter in (b) and (c) is not random*. Rather, that scatter results from the projection of many points from a smooth, undulating surface onto a single axis. We see also that doing a linear regression along SPL or f_o would not do justice to the shape of that surface, except perhaps within small parts of the range. In statistical terms, we see that the metric Q_Δ has both SPL and f_o as strong and non-linear covariates, which we will have to control for, if we are to detect the effects of our interventions reliably. It turns out that this applies to practically all metrics of the voice (for example, [13][14][15]).

Note that in Figure 6a, the vertical axis shows *the number of visited map cells*, not the number of phonated cycles. It does not say how many times Q_Δ fell in a given bin during the completion of the task, but rather how the metric was distributed across the elicited range of the voice. The beauty of this is that it does not matter if the participant phonated a lot more in some cells than in others – which would be very hard to avoid. The same applies to Figure 6b and 6c: each point is not the metric's value from one phonatory cycle, but the mean value of the metric from one cell in the map.[iv]

The localising principle

This brings us to the very important principle of *localisation*: a given voice tends to return to the

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

same value of a given metric whenever the same combination of f_o and SPL is revisited. ([17] pp 8-10). In consequence, voice maps are highly consistent *within individuals.* If an individual is asked to repeat the same task under the same conditions, the second map usually closely resembles the first [18]. However, if we ask someone else to do it, or if the same person's voice has changed in some way, intentionally or not, the map will not be so similar. While full demonstration cannot be provided here, examples are available in the supplementary data provided with [8] and [27]. Nevertheless, this observation supports the notion that the map reflects an individual's current vocal status and/or vocal behaviour. Individuality means that comparisons across interventions are meaningful within individuals, while comparisons to population norms will be less meaningful.

The localising principle implies that to cover as many cells as possible in a reasonable time, participants should be encouraged to change pitch and loudness continuously, moving around on the map and 'painting' as complete an area as necessary. The legacy method of sustaining vowels at constant pitch and loudness is here a waste of time: it will only give more replications of the same value, in the same small place on the map. While more replications often are better, we should not collect more than is needed. With voice mapping, some cells will inevitably contain data from many phonatory cycles while other cells contain only a few, but this is of little consequence. Tens of cycles or more per cell are generally sufficient to yield stable averages. A primary criterion for ''sufficient' cycles per cell is that the map should not appear grainy or speckled, and should have smooth colour gradients. Metrics related to entropy (irregularity), reflecting a certain degree of (voice) instability from cycle to cycle, typically require more cycles per cell for their mean values to stabilise than other metrics.

Can we trust the map?

In Figure 6d, each coloured pixel or 'cell' in the map was visited by at least six cycles, and there are 1172 qualifying cells. A minimum threshold of 6 cycles per cell was imposed to suppress outliers. The average number of visits to each cell was 76 cycles – more in the central parts, less on the edges. In total, the map contains data from 89439 cycles, so yes, we can trust it. It shows convincingly that the metric changes systematically with both SPL and f_o, and that those dependencies are not simply linear. Such multidimensional variation is typical of practically all metrics that we obtain from voice signals.

We must always be aware of possible shortcomings of our technical systems and how they might impact the appearance of the map. With some experience and reasoning, one can usually tell whether an unusual feature reflects a system artifact or a genuine property of the participant´s voice. For instance, if all maps show a systematic feature, such as a threshold that looks the same for all participants, then an issue with the technical system is the likely cause, since maps of different voices *should* be different.

Spacing the cells

We need to decide on a 'spatial' resolution, that is: how densely along the axes of f_o and SPL do we need to make the observations? The convention is a resolution of 1 dB in SPL and 1 semitone in f_o, according to a standard proposed in 1983 for phonetograms by the European Union of Phoniatricians [19]. So for making voice maps, we divide the voice field into cells that are 1 dB high and 1 semitone wide, and then sort the incoming metric data by the f_o and SPL at which they were obtained (Boxes 5 and 7).

7 RESOLUTION VERSUS 'EXPOSURE'

One decibel is close to the difference limen for auditory intensity, which is fine, while one semitone is somewhat coarse, by comparison. Still, we don't want the cells to be *too* small, because then the probability decreases that any given cell will be visited by at least one phonatory cycle, and we would have to make longer recordings. This is rather like a digital camera: smaller pixels (cells) give more detail, but also longer exposure times, since the number of photons (cycles) per second per pixel is finite. It turns out that 1 dB × 1 semitone is practical for most purposes.

To obtain detailed maps of a relevant range, such as that of habitual speech, we have to fill many of the cells, so we need our machinery to make many observations (long 'exposure'). On the other hand, in the interest of time, we want to make only as many observations as are needed – not more. As noted earlier, for most aspects of phonation the highest temporal resolution we can get is that of one phonatory cycle, which results in hundreds of observations per second. Typical frame lengths are rather longer than cycle lengths, and fixed. Therefore, a cycle-synchronous analysis is usually preferred over a frame-by-frame analysis.

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

Elicitation noise

When eliciting sustained vowels as “soft, comfortable, loud”, it is very unlikely that the participant will respond with exactly the same “comfortable pitch and loudness” on the occasions of pre- and post-intervention [20]. It has been shown that the *expected* discrepancy between occasions is on average ±5 dB in SPL, and may exceed ±10 dB; while for f_o, the expected discrepancy is ±3 semitones, with outliers up to an octave, for the same elicitation of the same task performed by the same participant. This variability introduces an observational error that we might call ‘elicitation noise’. To make things worse, the softer half of the habitual speech range tends to include those parts of the full voice range where gradients of the various metrics are at their steepest. In terms of communication efficiency, this makes good sense: we probably strive to speak in a range where small changes in input effort give large changes in modified output. This means, however, that the ‘elicitation noise’ is in fact concerningly large, and can obscure an intervention effect that one is looking for [14].

Difference maps

So how can we deal with elicitation noise? This is where mapping really comes into its own: let a participant make one voice map before an intervention, and then repeat the same task afterwards. In places where the islands on the two maps overlap, we can compute the differences cell by cell, and plot a new map of those differences. Such *difference maps* show what has changed across the intervention, in great detail (Figure 7).

The practical advantage of difference maps is that we do not have to steer the participant to repeat the task at certain precise levels and pitches. It suffices to get the participant to vocalise in at least some relevant region that overlaps pre- and post the intervention, for instance, by reading a text. If no overlap occurs because the voice has shifted substantially, then this shift itself is a primary effect of the intervention.

The theoretical advantage of difference maps is that they control for the metric’s covariates, SPL and f_o, regardless of how non-linear that covariation might be, thereby clarifying the residual effects of an intervention. Moreover, because difference maps are based on very many observations, from every phonated cycle, they are sensitive enough to reveal subtle changes. For instance, Engström *et al*. [22] investigated the effect of tilted body position on various voice metrics, including the EGG contact quotient Q_{ci}. This is of interest, because for magnetic resonance imaging of bodily functions, the subject has to lie down. An example from their results is shown in Figure 8.

Very often we find that the effect of an intervention varies in different parts of the voice range, even

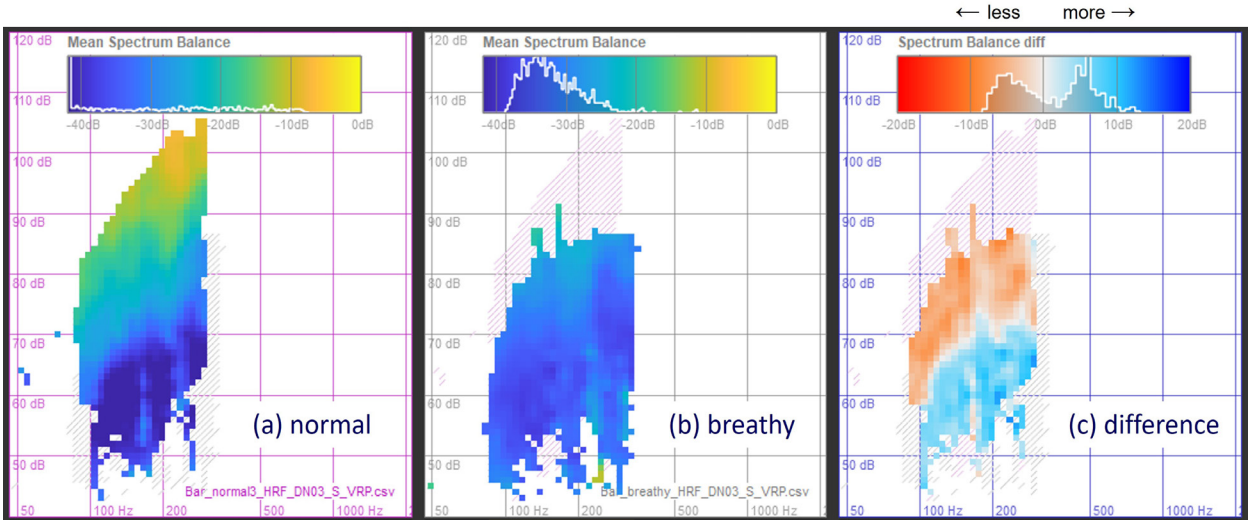


Figure 7. A trained baritone sang soft-loud-soft throughout all of his modal (chest) register. (a) Map of the spectrum balance SB, when the task was sung normally, (b) ditto, when the same task was intentionally sung in a very breathy voice. In the overlapping area only, the difference (b)-(a) can be computed cell by cell, as in (c), where red signifies a decrease in SB in breathy voice, while blue signifies an increase in SB. By comparing the SB layers shown here with the QΔ layers of these two maps, we find that the border between red and blue coincides with the onset of vocal fold contacting; that is, the change in SB was different depending on the type of phonation. The horizontal axis is fo in Hz, the vertical axis is SPL @ 30 cm in dB(C). Data adapted from [21].

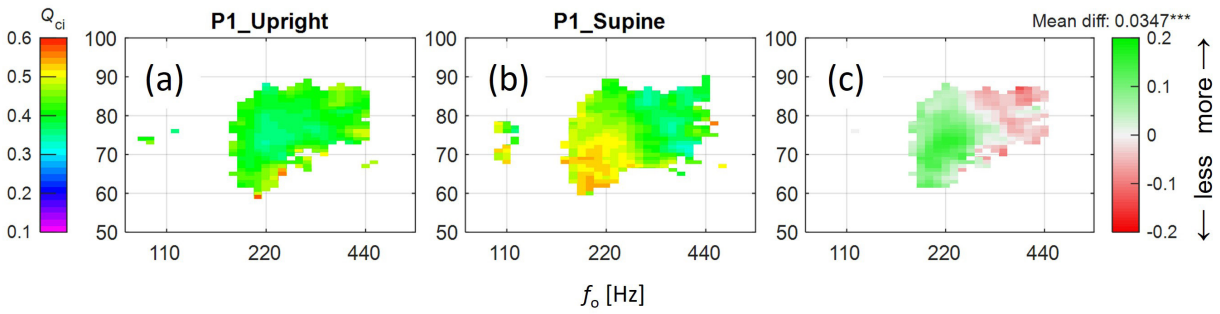


Figure 8. Example of detecting subtle changes in EGG contact quotient across conditions. The vertical axis is SPL in dB(C) at 0.3 m. An untrained adult female speaker read "The North Wind and the Sun" standing upright (a) and lying supine (b). This difference map (c) renders increase as green and decrease as red. It shows how the EGG contact quotient Q_{ci} increased when the speaker was supine, but only at the lower pitches. Here the "Mean diff" was computed from all the cells in the difference map, and so is very small. The horizontal axis is f_o in Hz, and the vertical axis is SPL@ 30 cm in dB(C). Data adapted from [22].

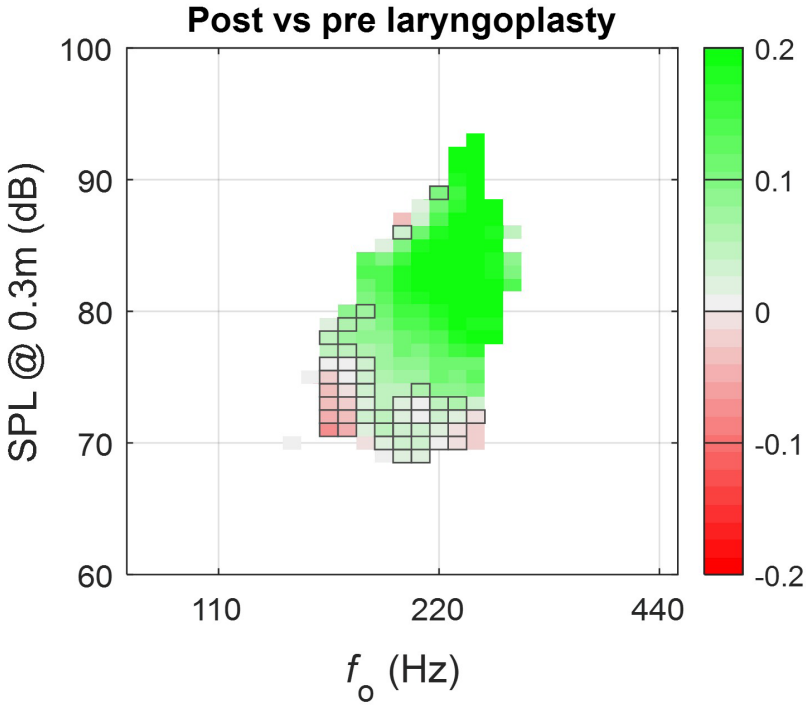


Figure 9. Difference map of the contact quotient Q_{ci} of a female laryngoplasty patient reading The North Wind and the Sun twice, pre- and post-intervention. Cells containing non-significant differences are marked with squares, 95% confidence level. For computing the statistics, only cells with at least 6 phonatory cycles have been included. In this case the laryngoplasty resulted in a markedly higher contact quotient, except in the softest range.

within the rather small range of habitual speech. This can be seen in Figure 8c: in the supine position, the Q_{ci} increased around 220 Hz, but it decreased at the higher pitches.

Are these differences 'significant'?

Null hypothesis significance testing of difference map outcomes can be challenging, since one first must specify which parts of the phonated range that should be considered in the test. Given the inherent variability of voice metrics, however, it is very difficult to make well-informed interpretations of the observed differences without a map. For example, the differences in Figure 7c, if averaged to a single value, would be close to zero, which is clearly an unsuitable value for characterising what we see there.

Since every cell in the map is based on observations of many cycles, one possibility is to do statistics cell-by-cell (Box 8: per-cell distributions). We can

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

compute a standard deviation and a confidence interval for the mean of every cell in the map. This enables us to determine a statistical confidence measure for every cell in the *difference* map, too. At a 95% confidence level, non-significant cells in the difference map can be marked as excluded, while the remainder are retained. Doing so shows that if there are contiguous and smooth regions of the same polarity (increase or decrease), with enough cycles (say, 6 or more per cell), then a 95% level of confidence in those regions is practically certain. Only the neutral-coloured cells (close to grey in Figure 8c) represent non-significant differences.

This approach has the disadvantage that it operates cell-by-cell, without considering the values of the surrounding cells, which leads to an overly conservative estimate of the statistical significance. Other ways of assessing the confidence in the observed differences would be desirable. In practice, conventional tests of 'statistical significance' are largely unnecessary here [23]. Clinicians, and increasingly also AI models, are trained to interpret images, and voice maps are just another type of image. However, we still have to decide *how large* the observed effects need to be, in order to be clinically relevant, for any given metric.

8 PER-CELL DISTRIBUTIONS

Distributions can be informative, even when considered for each individual cell in the map. Accumulating the number of cycles *N* falling into each cell yields a *density map*, which is in effect a 2D histogram of the occurrence (Figure 10a). Accumulating the period times of those cycles per cell, we get a map of the time spent in different parts of the voice range, which looks much the same, but is weighted more toward lower f_o. If we compute the mean of a given metric in the cell and display that, we get a map layer for that metric; this is the most common presentation. In principle, every cell can contain a distribution of observations, with an *N*, a mean, and a standard deviation. For instance, a map of the per-cell standard deviation of any EGG metric, rather than of the mean, typically reveals a distinct 'mountain ridge' along the threshold of contacting, where the variability of EGG metrics is greatest. Some distribution statistics are not meaningful until the entire data set has been seen, and they take some time to compute, so they are not convenient to display in real time. In offline post-processing, however, this is straightforward to do.

From individual metrics to phonation types

The search for a single universal metric of voice has been pursued for a long time. It is gradually emerging that the values of any individual metric are rarely, if ever, unambiguously relatable to the voice status of a participant. Though we may find it frustrating, after several decades of trying to measure the voice, the clinical community's informed faith in objective, quantitative measures has yet to be established. In particular, ascertaining consistent pathology thresholds for individual metrics has proven to be elusive.

Voice mapping can help us understand why this is so, by visualising the large intra- and inter-speaker variability observed in most metrics. It also facilitates a comparison of how different metrics describe the same voice production task. Figure 10[v] shows no less than nine metric layers (a)-(i) of a map of a highly trained baritone (same recording as in Figure 7a). He is very systematically sweeping from soft to loud to soft (*messa di voce*), advancing to a higher pitch, and repeating, while staying in modal voice. Since the singer was skilled, the recording took only 5½ minutes to make. Most of these metrics have some features in common, except for the Density (a). For instance, the EGG-derived metrics in (e) to (i) consistently have a clear transition from non-contacting to contacting, at about 65 dB SPL. This transition is especially clear in the peak dEGG (g). Yet it can also be seen that the metrics in Figure 10 all report on different aspects of phonation.

Across the voice range, the *role* of a given metric will often change. For instance, while an increased amount of jitter is expected in soft voice, it can also be seen in a loud falsetto, without necessarily being an indicator of suboptimal phonation. This brings us to the next higher level of representation: phonation or voice types, which is a concept that is familiar to clinicians. Specific types of phonation or types of voice typically exhibit specific combinations of metric values. Several research groups have devised so-called indexes that are statistically derived linear combinations of several metrics, such as the Acoustic Voice Quality Index (AVQI) [24], the Cepstral Spectral Index of Dysphonia (CSID) [25], or the Multidimensional Voice Program (MDVP) [26]. The computation of an index typically imposes a linear model on each metric to achieve the highest possible correlation with a target assessment. However, the variations in a metric are usually not linear. Also, each of the constituent metrics still varies across the voice range, so the problem of elicitation noise remains. So, what to do? Well, if we plot the changing value *of an index* on a voice map, we can get an idea of how much it varies. We can then choose, for instance, to take its value at a highly representative combination of SPL

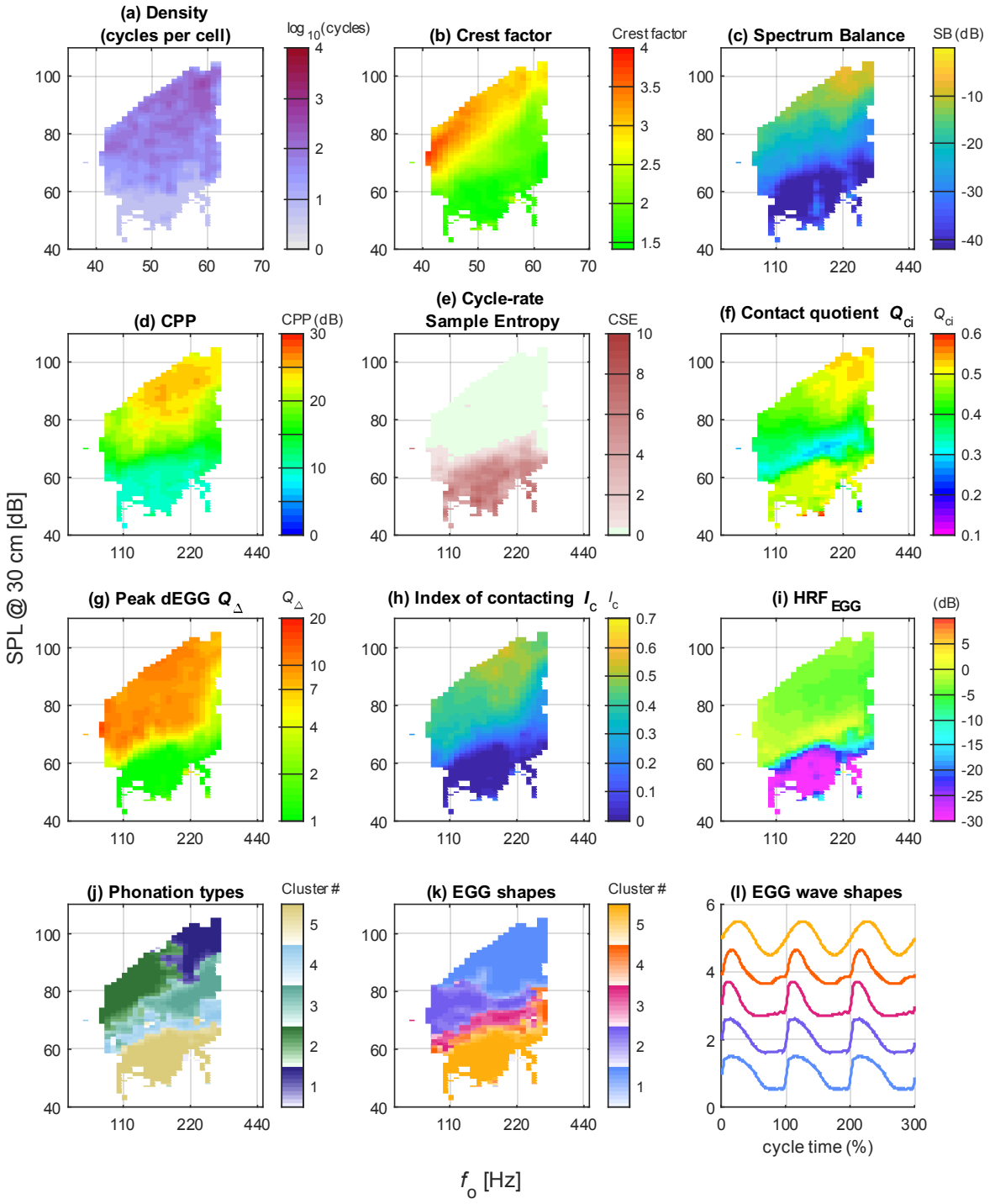


Figure 10. Multiple layers of a voice map from the same baritone production as in Figure 7a. (a) Density (cycles per cell) on a log scale from 1 to 10.000; (b) audio crest factor; (c) audio spectrum balance, expressed as the ratio of energy above 2 kHz to that below 1.5 kHz; (d) cepstral peak prominence (CPP); (e) EGG cycle-rate sample entropy (light green indicates highly stable EGG waveforms, brown indicates unstable waveforms); (f) EGG contact quotient (Q_{ci}); (g) normalised peak dEGG (Q_{Δ}); (h) index of contacting I_c (= Q_{ci} × log(Q_{Δ})) ; (i) the harmonic richness factor of the EGG; (j) five phonation type clusters found by k-means clustering of the six metrics (b) – (g); (k) five EGG waveshape clusters found by k-means clustering of the Fourier descriptors of the EGG cycles; and (l) the corresponding wave shapes, reconstructed and normalised in amplitude. Each cell represents the average of the number of cycles shown in (a). All layers (a)-(k) are generated by FonaDyn in real time, and subsequently smoothed and interpolated across occasional vacant cells.

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

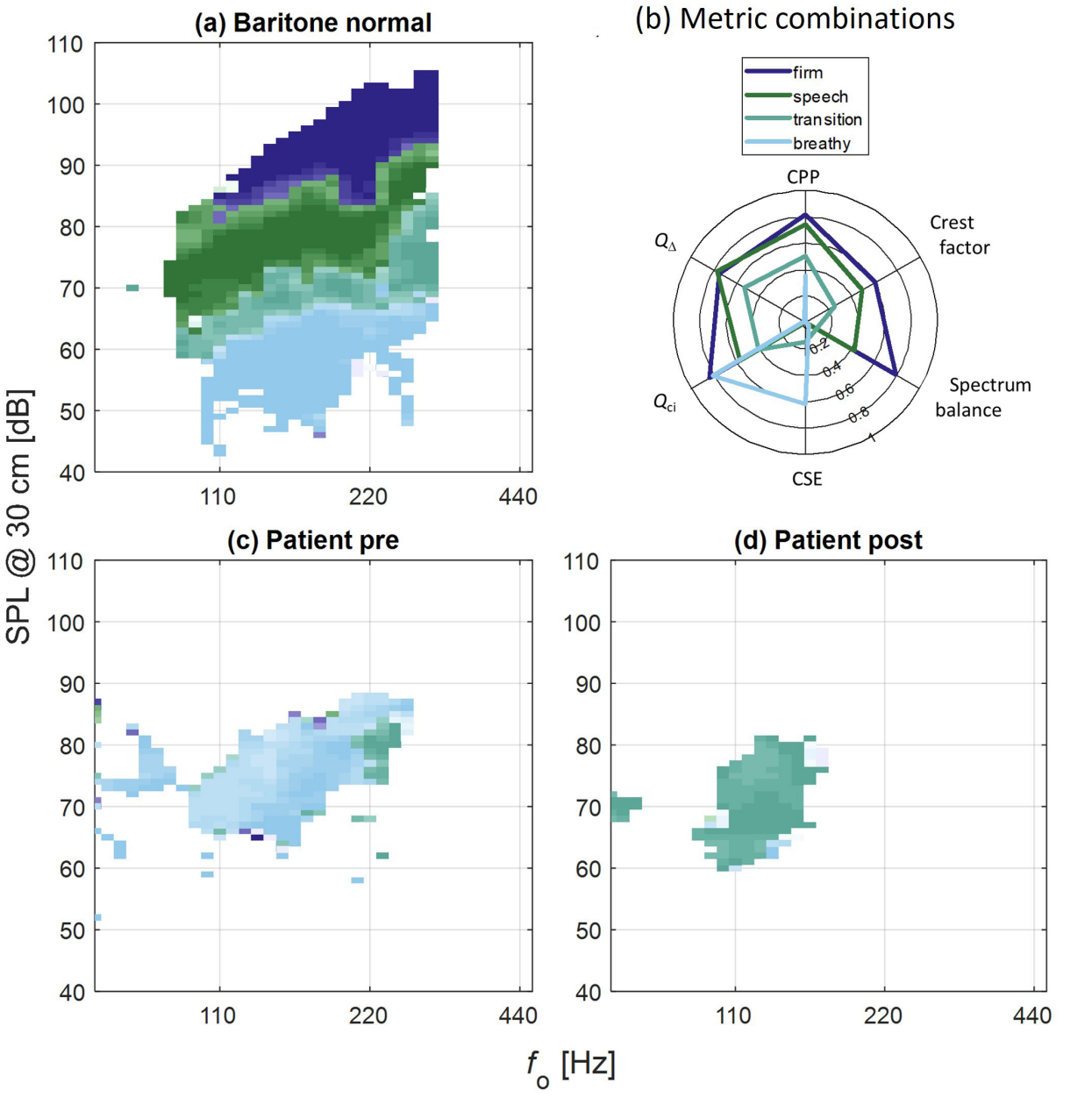


Figure 11. Example of phonation type clustering. (a) The 'trained baritone' map from Figure 10, classified into four manually selected phonation types; (b) the four corresponding combinations of six normalised voice metrics used to define phonation types for classification, with each metric scaled to a typical range of 0…1; (c) patient reading a text pre-laryngoplasty, the 'breathy' phonation type dominates; while in (d), post-laryngoplasty, the 'transition' phonation type dominates, indicating some voice quality improvement after the intervention, rather than merely a change in range.

and f_o, such as the location of the highest density in a map of a person's running speech [27].

Connecting to perception, via clustering

An index generates a single value from several input values. Another possibility is to examine how several input values tend to cluster into certain combinations. Due to the way the voice works, only some combinations of metric values are possible. By combining multiple metrics, we can come much closer to characterising the *types* of phonation. For example (Figure 11), a healthy soft breathy voice without vocal fold contacting is characterised by low spectrum balance, low CPP, a Q_{ci} of 0.5 [12], a Q_{Δ} close to 1, and high cycle-rate sample entropy (CSE). In contrast, healthy, loud, pressed voice typically exhibits a high spectrum balance, a high CPP, a Q_{ci} > 0.5, a high Q_{Δ} > 5 and a CSE of zero.

Fortunately, computers are really good at sifting through lots of data and finding the combinations of metric values that occur especially often; this is called 'statistical clustering'. And, unlike the computation of indexes, clustering does not impose a linear model, or any predefined model at all, and thus avoids the constraining assumption that each contributing dependency extends linearly across most of the voice range.

We can let the computer apply statistical clustering to a set of metrics that are thought *a priori* to be relevant [28]. Alternatively, the classification may be performed by expert listeners, interactively picking sounds straight from different places in the voice map [21] and storing those local metric combinations as classifiers. The clinician (or better, a consensus group of clinicians), can define their own subjective categories from representative samples of different phonation types, such as 'normal', 'breathy' or 'pressed', or even of whatever is relevant for a given client's condition.

In Figure 11a, the map of the trained baritone's modal voice has been partitioned by a clinician. By listening to vocalisations at different points in the map of Figure 10, she selected and labelled four phonation categories as 'firm', 'speech', 'transition' and 'breathy'. Figure 11b illustrates the relative values of six metrics for each of the categories; each such combination is called a 'centroid'. Rebuilding the map with these categories automatically resulted in Figure 11a. The same centroids were then applied to classify the productions of a voice patient with idiopathic vocal fold paralysis. The patient's task was to read twice *The North Wind and the Sun*, before and after injection laryngoplasty. Prior to the intervention (Figure 11c), the phonated region was dominated by the 'breathy' type, with hardly any vocal fold contact at all, whereas after the intervention (Figure 11d) the phonated region, although smaller, was dominated by the 'transition' type, with a limited amount of vocal fold contact throughout. This substantiates a change in phonation post-intervention. Although changes consistent with such a change were indeed seen in the maps of the individual metrics, their combined effects are more conveniently and intuitively summarised in a map of clusters.

On comparing Figure 11a (four clusters, manually specified) with Figure 10j (five clusters, derived automatically), we see that the loudest region then splits into three parts rather than two. Using all EGG metrics, and also three acoustic metrics, Cai and Ternström [28] found that clustering untrained normophonic voices into only two categories will reliably distinguish two zones across speakers, above and below the vocal fold contact transition; that is, breathy and normal voice. This would not have been apparent without EGG-based metrics. They reported also that little information was gained by going beyond six clusters in male voices, or beyond four clusters in female and child voices.

Another example of clustering is shown in Figure 10, (k) and (l), where the algorithm identified five characteristic wave-shapes in the EGG signal only. The corresponding map regions of EGG wave-shapes are only partially similar to the phonation type regions shown in Figure 10j.

The metrics in Figure 10 should not be construed as a conclusive or comprehensive set. For instance, while work so far has concerned mainly phonation, with the EGG signal as a complement to the acoustic signal, metrics of *articulation* (vocal tract shape and size) have not yet been addressed. Such metrics would be related to the resonance properties of a voice, which would be relevant for training articulation in singing or for quantifying gender-affirming voice changes across interventions. It is also possible to acquire other physiological signals in parallel, and to relate them to voice production by mapping. For instance, in study [29], per-cell *correlations* between the lung volume, larynx height and contact quotient were mapped in female singers. These correlations also varied both across the voice range, and between individuals. Some singers consistently lowered their larynx at high lung volume, while others did it only in certain parts of the range of the song.

Finally, it should be possible to quantify the so-called 'mutual information' from the map layers of multiple metrics, so that we might omit those metrics that contribute too little, for a mainstream clinical context. Agreement on a core inventory of particularly informative metrics could allow standardisation of characteristic colour scales for each. The various colour scales used in Figure 10 should therefore be seen as illustrative examples rather than definitive standards.

Concluding remarks

A major benefit of voice mapping, impossible to do justice in writing, lies in the real-time feedback it provides, for both the client and the therapist. 'Drawing with your voice' leaves an immediate and concrete trace of what one's voice can do, and how that compares with earlier efforts. This is engaging and interesting, offering valuable insights to client and therapist alike [30]. Voice mapping is applicable not only to sustained phonation but also to running speech, and, to varying extents, with most pathological voices encountered so far. Even though signals of severely pathological voices often have weak periodicity, there are usually at least some regions in the voice field where the mapping gives interpretable results.

Communication • DOI: 10.2478/ijvs-2025-0002 IJVS • 1(1) • 2025 • 1–20

The primary concern of clinicians and voice pedagogues is not the values of individual metrics, but whether or not a voice is working as desired. So, higher-level representations of vocal function need to be developed in close collaboration between voice scientists and practitioners, while remaining firmly grounded in physical measurement. The mapping of clustered metrics is one step in that direction. Although much of the methodology remains to be devised, agreed upon, and implemented, voice maps already offer two immediate benefits: (1) they visualise how voices are very variable and individual, and (2) they account for the considerable covariation of most metrics with f_o and SPL, allowing us to make maps of changes across interventions that are not obscured by effects of f_o and SPL alone.

Voice mapping, as described here, is not a universal solution to the problem of describing voices. For one thing, it completely disregards the *temporal* evolution of voice properties. Other mapping approaches exist, such as the Göttingen Hoarseness diagram [31], which aims to discriminate between pathologies.

We are currently in a time where the excitement around rapid developments in machine learning and AI is spawning a deluge of work on 'black box' algorithms for classifying voices, for example by pathology. For a critical discussion of this situation, see *Gómez-Vilda et al*. [32] . Since the voice map representation condenses millions of signal sample values into about one thousand data points per metric, it can serve also as pre-processed input for machine-learning algorithms, speeding up their training by orders of magnitude. Hopefully, new developments in data analytics will help us to tackle outstanding problems, such as inversion from acoustics to physiology, and the charting of how the voice is controlled in detail. The future looks promising.

Implementations

The examples shown in this article were made using FonaDyn, which can be freely downloaded for Windows and MacOS, with detailed documentation, on author S.T.'s profile page at https://www.kth.se/profile/stern. Alternative tools include *RecVox* by Svante Granqvist (www.tolvan.com, Windows only, freeware) which is smaller and has a proven clinical track record; and *Voice Profiler* by author P.P., of which a new free version is currently under development. An interest-group forum dedicated to voice mapping [33] was launched in 2025.

Acknowledgments

This overview article builds on work by the authors over many years, reflected in the number of self-citations in the references, and supported by several grants and institutions. The Royal Conservatory of The Hague (NL), although not a research institute, supported the longitudinal monitoring and documentation of its singing students. We acknowledge especially the KTH faculty grants for author S.T. that have enabled him to do extensive work on voice mapping. Additional sources of funding included the WROCAH programme (UK), the Wenner-Gren Foundations (SE), the Swedish Research Council, and the China Scholarship Council (CN). We are very grateful to Silvia Capobianco and Gunnar Björck for clinical collaborations on some of the examples mentioned here. Meike Brockmann-Bauser and anonymous reviewers offered insightful comments on the manuscript.

Declaration of conflict of interests

Author P.P. is the creator of the Voice Profiler system, and the sole proprietor of Voice Quality Systems, in which all commercial activities have ended as of June, 2017. Author S.T. is the creator of FonaDyn, in which he has no commercial interests.

Endnotes

i The time *derivative* of a signal is its rate-of-change; how much the amplitude changes per unit of time.

ii $QΔ$ assumes values close to a minimum of 1 when no vocal fold contact occurs, and values greater than approximately 2 when contact is present. Formally, it is defined as the cycle-normalised peak of the EGG time derivative [12].

iii The graphs in Figure 6 (a)-(c) were generated in a spreadsheet app, from the corresponding .csv file named at lower right. That .csv file, containing this voice map, was made and displayed using the *FonaDyn* software [16]. FonaDyn maps more than 20 metrics, each onto a separate map layer, in real time during phonation.

iv This is why vertical striations are visible in Figure 6 (b) and (c): the values on the abscissa (the x-axis) are rounded to whole semitones or whole decibels. The map in (d) is not a 2D histogram, since the colouring indicates the average value of a metric rather than the occurrence of observations within a cell.

v The various colour scales in Figure 10 are configurable by the FonaDyn user. They are shown here as examples only, and are not intended to be prescriptive. Ideally, recommended colour scales for different metrics should be established through international expert consensus.

References

[1] Baken, R. J. (1992). Electroglottography. *Journal of Voice, 6*(2), 98–110. https://doi.org/10.1016/S0892-1997(05)80123-7

[2] Herbst, C. T. (2020). Electroglottography – An Update. *Journal of Voice*, *34*(4), 503–526. https://doi.org/10.1016/j.jvoice.2018.12.014

[3] Chi, Y., Honda, K., & Wei, J. (2022). Near-infrared photoglottography for measuring multiple glottal events. *JASA Express Letters*, *2*(10). https://doi.org/10.1121/10.0014810

[4] Zañartu, M., Ho, J. C., Mehta, D. D., Hillman, R. E., & Wodicka, G. R. (2013). Subglottal Impedance-Based Inverse Filtering of Voiced Sounds Using Neck Surface Acceleration. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(9), 1929–1939. https://doi.org/10.1109/TASL.2013.2263138

[5] Selamtzis, A., & Ternström, S. (2014). Analysis of vibratory states in phonation using spectral features of the electroglottographic signal. *The Journal of the Acoustical Society of America*, *136*(5), 2773–2783. https://doi.org/10.1121/1.4896466

[6] Chen, C. J. (2016). *Elements of Human Voice*. World Scientific. https://doi.org/10.1142/9891

[7] Ternström, S. (2005). Does the acoustic waveform mirror the voice? *Logopedics Phoniatrics Vocology*, *30*(3–4), 100–107. https://doi.org/10.1080/14015430500238400

[8] Pabon, P., & Ternström, S. (2020). Feature Maps of the Acoustic Spectrum of the Voice. *Journal of Voice*, *34*(1), 161.e1-161.e26. https://doi.org/10.1016/j.jvoice.2018.08.014

[9] Ternström, S., Bohman, M., & Södersten, M. (2006). Loud speech over noise: Some spectral attributes, with gender differences. *The Journal of the Acoustical Society of America*, *119*(3), 1648–1665. https://doi.org/10.1121/1.2161435

[10] Ternström, S., Pabon, P., & Södersten, M. (2016). The voice range profile: Its function, applications, pitfalls and potential. *Acta Acustica United with Acustica*, *102*(2), 268–283. https://doi.org/10.3813/AAA.918943

[11] Pabon, J. P. H., & Plomp, R. (1988). Automatic phonetogram recording supplemented with acoustical voice-quality parameters. *Journal of Speech and Hearing Research*, *31*(4), 710–722. https://doi.org/10.1044/jshr.3104.710

[12] Ternström, S. (2019). Normalised time-domain parameters for electroglottographic waveforms. *The Journal of the Acoustical Society of America*, *146*(1), EL65–EL70. https://doi.org/10.1121/1.5117174

[13] Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C. (1994). Relationships Between Intra-Speaker Variation in Aerodynamic Measures of Voice Production and Variation in SPL Across Repeated Recordings. *Journal of Speech, Language, and Hearing Research*, *37*(3), 484–495. https://doi.org/10.1044/jshr.3703.484

[14] Iob, N. A., He, L., Ternström, S., Cai, H., & Brockmann-Bauser, M. (2024). Effects of Speech Characteristics on Electroglottographic and Instrumental Acoustic Voice Analysis Metrics in Women With Structural Dysphonia Before and After Treatment. *Journal of Speech, Language, and Hearing Research: JSLHR*, *67*(6), 1160-1681. https://doi.org/10.1044/2024_JSLHR-23-00253

[15] Brockmann-Bauser, M., Bohlender, J. E., & Mehta, D. D. (2018). Acoustic Perturbation Measures Improve with Increasing Vocal Intensity in Individuals With and Without Voice Disorders. *Journal of Voice*, *32*(2), 162–168. https://doi.org/10.1016/j.jvoice.2017.04.008

[16] Ternström, S. (2024). Update 3.1 to FonaDyn – a system for real-time analysis of the electroglottogram, over the voice range. *SoftwareX*, *26*, 101653. https://doi.org/10.1016/j.softx.2024.101653

[17] Pabon, P. (2018). Mapping individual voice quality over the voice range: The measurement paradigm of the voice range profile [Doctoral dissertation, KTH Royal Institute of Technology]. KTH DiVA portal. https://kth.diva-portal.org/smash/get/diva2:1253755/FULLTEXT01.pdf

[18] Ternström, S., & Pabon, P. (2022). Voice Maps as a Tool for Understanding and Dealing with Variability in the Voice. *Applied Sciences (Switzerland)*, *12*(22), 11353. https://doi.org/10.3390/app122211353

[19] Schutte, H. K., & Seidner, W. (1983). Recommendation by the Union of European Phoniatricians (UEP): Standardizing voice area measurement/phonetography. *Folia Phoniatrica et Logopaedica*, *35*(6), 286–288. https://doi.org/10.1159/000265703

[20] Brown, W. S., Murry, T., & Hughes, D. (1976). Comfortable effort level: An experimental variable. *Journal of the Acoustical Society of America*, *60*(3), 696–699. https://doi.org/10.1121/1.381141

[21] Jacobsson, L. (2023). Listeners' perception of voice quality: Perceptual and acoustic analyses. [Unpublished student project report]. Department of Neuroscience and Physiology, Division of Logopedics, University of Gothenburg.

[22] Engström, H., Włodarczak, M., & Ternström, S. (2024). Mapping the effect of body position: Voice quality differences in connected speech. In M. Heldner, M. Włodarczak, C. Ericsdotter Nordgren, & C. Wikse Barrow (Eds.), *Proceedings of FONETIK 2024, Stockholm, June 3–5, 2024:* 21–26. Stockholm University. https://doi.org/10.5281/zenodo.11396054

[23] Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". *The American Statistician*, 73(sup1), 1-19, https://doi.org/10.1080/00031305.2019.1583913

[24] Maryn, Y., de Bodt, M., & Roy, N. (2010). The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders. *Journal of Communication Disorders*, *43*(3), 161–174. https://doi.org/10.1016/j.jcomdis.2009.12.004

[25] Lee, J. M., Roy, N., Peterson, E., & Merrill, R. M. (2018). Comparison of Two Multiparameter Acoustic Indices of Dysphonia Severity: The Acoustic Voice Quality Index and Cepstral Spectral Index of Dysphonia. *Journal of Voice*, *32*(4), 515.e1-515.e13. https://doi.org/10.1016/j.jvoice.2017.06.012

[26] Deliyski, D. D. (1993). Acoustic Model and Evaluation of Pathological Voice Production. *3rd European Conference on Speech Communication and Technology, EUROSPEECH 1993*, 1969–1972. https://doi.org/10.21437/eurospeech.1993-445

[27] Patel, R. R., & Ternström, S. (2021). Quantitative and qualitative electroglottographic wave shape differences in children and adults using voice map–based analysis. *Journal of Speech, Language, and Hearing Research*, *64*(8), 2977–2995. https://doi.org/10.1044/2021_JSLHR-20-00717

[28] Cai, H., & Ternström, S. (2022). Mapping Phonation Types by Clustering of Multiple Metrics. *Applied Sciences (Switzerland)*, *12*(23), 12092. https://doi.org/10.3390/app122312092

[29] Ternström, S., D'Amario, S., & Selamtzis, A. (2018). Effects of the Lung Volume on the Electroglottographic Waveform in Trained Female Singers. *Journal of Voice*, *34*(3), 485.e1-485.e21. https://doi.org/10.1016/j.jvoice.2018.09.006

[30] Holmberg, E., Ihre, E., & Södersten, M. (2007). Phonetograms as a tool in the voice clinic: Changes across voice therapy for patients with vocal fatigue. *Logopedics Phoniatrics Vocology*, *32*(3), 113–127. https://doi.org/10.1080/14015430701305685

[31] Fröhlich, M., Michaelis, D., Strube, H. W., & Kruse, E. (2000). Acoustic Voice Analysis by Means of the Hoarseness Diagram. *Journal of Speech, Language, and Hearing Research*, *43*(3), 706–720. https://doi.org/10.1044/jslhr.4303.706

[32] Gómez-Vilda, P., Gómez-Rodellar, A., Palacios-Alonso, D., Rodellar-Biarge, V., & Álvarez-Marquina, A. (2022). The Role of Data Analytics in the Assessment of Pathological Speech—A Critical Appraisal. *Applied Sciences*, *12*(21), 11095. https://doi.org/10.3390/app122111095

[33] Ternström, S. (2025, August). Voice Mapping Central [Online forum]. https://voicemapping.groups.io/