Analysis of the electroglottogram over the voice range – the FonaDyn system

Abstract

Background and objective: the human voice production mechanism has many degrees of freedom, making it difficult to quantify and assess vocal status from the acoustic signal alone. Combining acoustics with electroglottography could help account for the challenging variability that is encountered over the soft-loud and low-high ranges of the voice.
Method: the shapes of electroglottographic pulses are characterized by how their Fourier descriptors cluster in a high-dimensional space. The clusters are visualized by mapping them onto the frequency-level plane of the so-called voice range profile. The sample entropy of the Fourier descriptors is used similarly to map phonatory transitions and instabilities. The system has been applied in several voice analysis and training scenarios.
Results and conclusion: We find that, when the cluster-learning phase is tailored specifically to the user's research question, the method is able to classify and/or stratify various phonatory regimes of interest, in real time, with visual feedback. Its novel contributions are: phonatory regimes and voice instabilities are mapped over voice sound level and phonation frequency; statistical clustering obviates the need for pre-defining thresholds or categories; and, the sample entropy shows promise as a metric of perceptually relevant voice instabilities. FonaDyn is hereby provided to the voice research community, as freeware under public license.

Keywords:
electroglottography; voice range profile; voice analysis; phonation; cluster analysis; sample entropy; SuperCollider

1 Introduction

The human voice produces its sounds using a very complicated system of biomechanics, aerodynamics and acoustics, with very many degrees of freedom, all under exquisite control from the brain. The larynx with its vibrating vocal folds is central to voice production, and thus of great clinical and pedagogical interest. But it is also sensitive, and not readily accessible for non-invasive inspection and measurement. Even the sound that emerges at the lips is only indirectly connected to the vocal fold vibration, being subject also to the highly variable acoustic transfer function of the vocal tract, from larynx to lips.

With current technology, one of the few signals that is directly related to vocal fold vibration, and that can be obtained non-invasively and at low cost, is the electroglottogram (EGG) [1]. The electroglottograph provides an electrical signal, the alternating part of which is proportional to the instantaneous area of contact between the colliding vocal folds. While the EGG signal is well known to vocologists, the dynamic variation of its features over the voice range have not been extensively researched, nor has it been modelled in any great detail. A system that can classify EGG pulse shapes and map them across low-high and soft-loud phonation can be expected to give important insights into normal and pathological voice in speech and in singing.

This article describes a new real-time EGG analysis system that is based entirely on open-source free software and readily available medium-cost hardware. The system is called FonaDyn (for ‘phonatory dynamics’). It has been pre-researched [2][3][4][5] and used in pilot studies [6].

1.1 Background

The operating limits of a person’s voice range are of clinical and pedagogical interest, and are often documented using a so-called voice range profile (VRP) [7]. However, the limits of the voice depend on many things, as do the properties of the acoustic signal. A quest for a better account of the many sources of variation that influence the VRP began within a research project on phonatory dynamics and states, conducted at KTH in 2011-2013. It has since generated further investigations, and has prompted the development of the FonaDyn system.

One source of variation in the voice is the ability of the vocal folds to vibrate in several different ways. This relates to the concept of vocal ‘registers’, or ‘phonatory mechanisms’ [8]; the two most often discussed mechanisms being modal/chest/M1 voice, as contrasted with falsetto/head/M2 voice, which are of particular interest in singing, but also in clinical settings. Selamtzis and Ternström [3] explored the EGG for discriminating automatically between the M1 and M2 modes of phonation, and found that this is possible to a useful level of confidence. They used an offline procedure in Matlab® [9], which has informed the design of the present system.

Other examples of phonatory dimensions that can be studied with FonaDyn include vocal loudness [4] and degree of vocal fold adduction [6]. Although the system was designed with the EGG signal in mind, it can be used also with any other periodic signal derived from phonation, such as the photoglottogram (for glottal area) or a signal from a neck-mounted accelerometer.

Other efforts to characterize the EGG signal have focused mainly on deriving various scalar metrics for the pulse shape. This necessitates the definition of time-domain thresholds, which can be problematic [10]. One of few studies to account for the shape of the whole EGG pulse did so using principal component analysis [11]. Herbst *et al*. proposed the wavegram, a rich visualization of the EGG [12], which however defers the classification task to a human observer.

1.2 Design considerations

For the development of FonaDyn, the following principal goals were defined.

- For real-time visual feedback in the clinic or singing studio, the system should operate with a perceptually negligible delay from sound to display,
- the great majority of all glottal cycles should be correctly detected and analyzed without gaps in the data, for phonation frequencies of at least 100 – 1000 Hz,
- the user interface should be adequate for research purposes, but not for a commercial product,
- the resulting data should be easy to export to other software tools such as Matlab or spreadsheets, for further visualization and analysis,
- the system should be based on open-source, cross-platform tools.

The success of the system will be judged by the new insights it affords into the mechanisms of voice production, including the diagnosis of some voice problems or challenges; and possibly by its potential as a tool for visual feedback in vocal training and/or therapy.

2 Computational methods and calculation

2.1 Purpose

We seek to map regions of different vocal fold vibratory regimes, across different vocal effort levels and different phonation frequencies. The vocal folds are three-dimensional structures, and so a complete characterization would need more than a one-dimensional signal. The EGG gives us only a relative and uncalibrated measure of the instantaneous total area of contact between the medial surfaces of the vocal folds. This is like trying to study the motion of a contorting, three-dimensional object by observing only the brightness of a silhouette image of the object – which of course is far from ideal. Still, the range of plausible positions and shapes of the vocal folds is rather well known, constraining the possible interpretations [13][14]. Especially when combined with (experience of) other modes of observations, the EGG pulse waveform can provide very useful information on the type of phonation.

The motion of the vocal folds in phonation is essentially periodic, and they can vibrate in several vibratory modes. We therefore resort to Fourier analysis of the EGG signal, on the assumption that especially its first few Fourier descriptors (their magnitudes and phases) will be related to the vibratory contact pattern, albeit in an indirect manner. K-means cluster analysis is used to discriminate between patterns dominated by one mode or another; or, to stratify the EGG pulse shapes over some parameter continuum. We are interested also in detecting transitions between modes (i.e., 'voice breaks') and other instabilities. For this we adopt the so-called sample entropy, applied to the Fourier descriptor data.

In the rest of this section, we describe the design choices and the resulting processing steps in some detail. An overview of the major signal paths is given in Figure 1.

2.2 Recording and playback

In FonaDyn, analysis can be made of the live incoming voice and EGG signals, or of the same signals replayed from file. For pedagogical work or voice training, real-time visual feedback is given of the live signals and of the analysis outcomes. For experimental work, recordings to file are typically made for later analysis. When recording, the raw input signals are written to a two-channel WAV file.

The system sample rate is 44100 Hz. Voice+EGG files acquired on other systems can be analyzed, and many different soundfile formats can be read, as long as they are 2-channel files sampled at 44100 Hz per channel. Even when analyzing from a pre-recorded file, the program has to run in real time; it cannot process silently at a higher speed.

2.3 Preconditioning

The analog electronics preceding the A/D-converter of the digital audio interface will have a DC blocker and the compulsory anti-aliasing filter, but should apply no other filtering. On input to FonaDyn, the digitized EGG signal is first preconditioned numerically (high-pass at 100 Hz) so as to suppress the low-frequency heaving that is typical of some EGG hardware, and to reduce high-frequency noise (steep low-pass at 10 kHz). These filters are phase-linear finite impulse response (FIR) filters. Their presence means that the lowest phonation frequency that can be correctly analyzed is 100 Hz, while the highest is 10 kHz divided by the number of DFT components chosen for the analysis. Exceeding these limits is possible, but will introduce a possibly undesired dependency on phonation frequency.

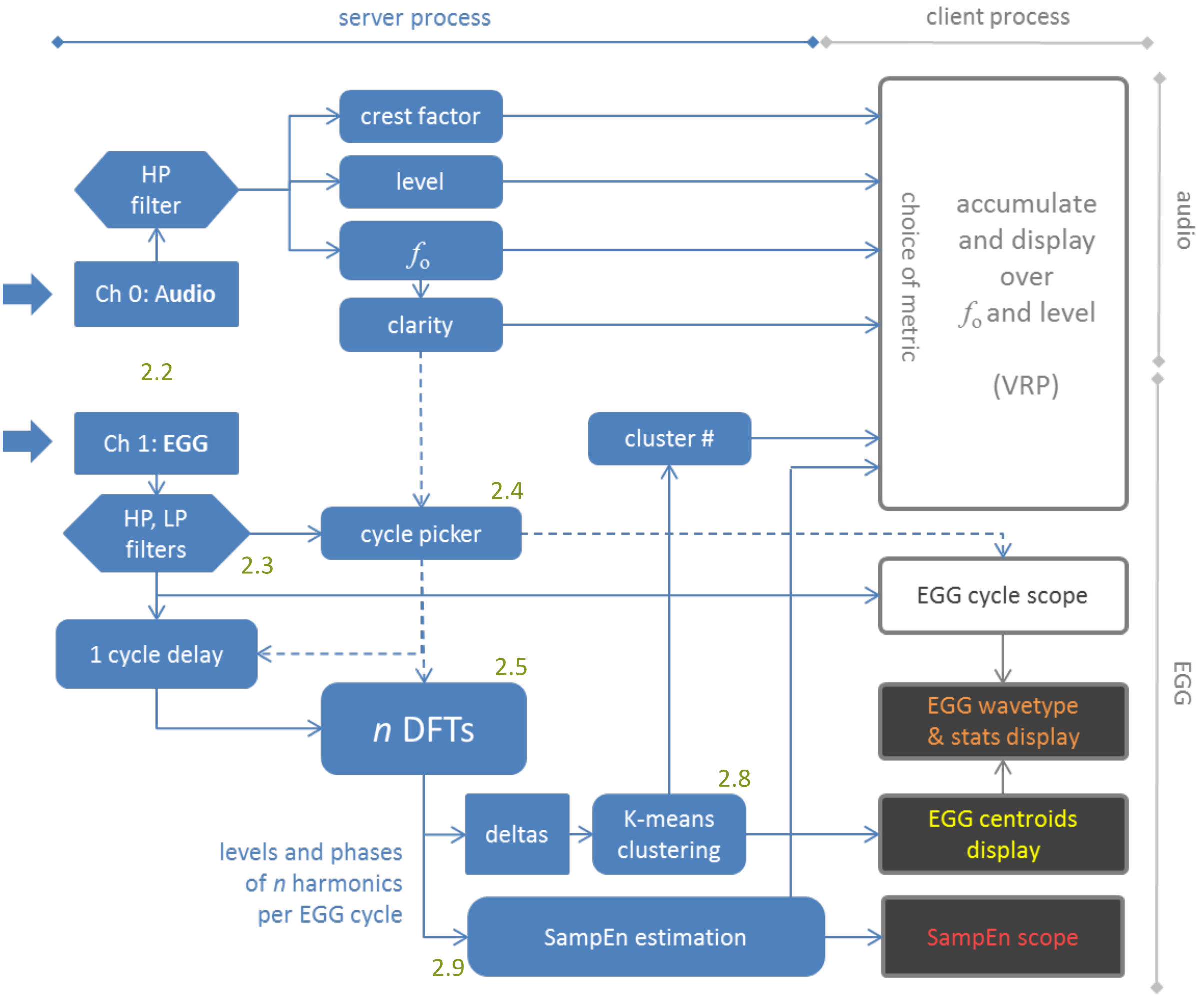


Figure 1. Functional block diagram of the signal paths in FonaDyn. Numbers refer to the text sections that follow. For clarity, some detail has been omitted.

2.4 Cycle separation

The second step is to identify and mark each glottal cycle in the EGG signal. Segmenting the individual periods of a quasi-periodic waveform is a classical problem in speech analysis. For the EGG signal, which has a simpler waveform and a more uniform spectrum than most speech sounds, this problem is somewhat easier. Still, it is not trivial. Here, the EGG cycles are found using a phase tracker. The sampled input signal $E(t)$ is passed through a leaky integrator, yielding a new signal $\int E(t)dt$, or $\int E$ for short. The integrator introduces a phase shift of $\pi/2$ radians at all frequencies, and also attenuates noise. If E and $\int E$ are both bipolar signals, then plotting E against $\int E$ yields a closed-loop pattern through all four quadrants, known as the *phase portrait*. The instantaneous phase φ_E of E is obtained as $\varphi_E = \arctan(E/\int E)$ radians. The phase φ_E is not necessarily monotonic over a cycle, so one or more loops can occur in the phase portrait. To minimize the impact of such loops, the signal representing φ_E is passed through a double-sided peak follower *ad modum* Dolanský [15], which generates trigger pulses each time the phase jumps from π back to $-\pi$. These pulses are used to mark the end of each EGG cycle.

Alternatively, the time-differentiated EGG signal (dEGG) can be computed and passed directly to the peak follower. This can work better for some particularly spiky EGG waveforms, while the phase tracker method works better in most other cases, especially for low-amplitude (pre-contacting) EGG waveforms. The phase tracker method is therefore the default in FonaDyn.

For each completed EGG cycle, the cycle detector produces a timing marker in the form of a trigger pulse. It computes also the length of the finished cycle, in sample points. For a tone phonated at about 200 Hz, for example, the waveform in each period is described by 44100/200 ≈ 220 sample points.

In parallel with this EGG cycle detection, the audio signal from the microphone is processed by a correlation-based pitch detector [16]. The pitch detector computes the fundamental frequency of the audio, and also a ‘clarity’ index that reflects the regularity of the audio signal. When the clarity index drops below a specified threshold, the corresponding EGG cycles are discarded. No further analysis is attempted of phonation that is too aperiodic, according to this criterion.

2.5 Spectrum analysis

For all cycles that are not discarded, the first N Fourier components are computed ($N = 2 \ldots 20$). Because the time window is chosen as one period of the EGG signal, this produces value pairs of magnitude and phase for the harmonic components 1, 2, …. N . A small quantization error arises here, because EGG period times are represented only as integer multiples of the sampling interval. In practice, though, this error is absorbed by the subsequent statistical clustering.

The Fast Fourier Transform (FFT) is not used; being constrained to analysis window lengths of integer powers of two, it cannot be cycle-synchronous in the general case. Instead, the Discrete Fourier Transform (DFT) components, or Fourier Descriptors (FDs), are computed directly, in the time domain. The DFT computational load is then constant in time (shorter cycles compute more quickly), and is still reasonable, because N is small. The EGG signal is delayed until a full cycle has completed; then, the DFT is applied to the waveform between the trigger pulses. Because the signal is already segmented into periods, the analysis weighting window can be rectangular (= no weighting).

2.6 Rationales

(A) EGG devices are difficult to calibrate for true contact area, for two reasons: (1) skin/electrode contact and tissue conductance will vary between subjects, and (2) the amplitude of the EGG signal changes as the larynx moves up and down in the neck, due to the changing distance between the electrodes and the vocal folds. Therefore we seek a method of characterizing the EGG waveform that disregards its absolute amplitude.

(B) The cycle separation procedure produces trigger pulses, as described above. The phase of the trigger pulse, relative to the phase of the fundamental of the EGG waveform, will inevitably change somewhat with the waveform shape; for instance, depending on whether the latter is nearly sinusoidal or has a steep slope at vocal fold collision. Therefore, we seek a method of characterizing the EGG waveform that disregards the varying phase of the cycle trigger pulse.

(C) The optimal choice of the number N of FDs to be computed may depend on the particular aspect of the EGG waveform that we wish to analyze. For small values of N, most of the high-frequency energy of the EGG waveform will be discarded. However, the steepness of closure is a very important aspect of the EGG, and it is closely related to high-frequency energy. It is therefore desirable to retain an estimate of that energy, even when N is small.

(D) The phase as obtained using the arctan function is bounded to the interval [-π, π [; should it increase beyond half a period, it will wrap around, causing a 2π discontinuity. This will unduly shift the corresponding cluster centroid and can potentially cause the clustering algorithm to spawn a new, disjunct cluster that does not represent a significantly different waveform. To avoid this, we seek a representation of the phases that is free from such discontinuities.

(E) For visual inspection, we wish to be able to resynthesize the 'average' EGG waveforms, as represented by each of the K-means clusters.

2.7 Choice of features

The features chosen for clustering are the *relative* levels and phases of the FDs $2 \ldots N$, using the level and the phase of the first component (the 'fundamental') as the reference. This answers to rationales (A) and (B) above. Furthermore, the residual 'power' in the EGG signal that is not accounted for by the first N FDs is estimated, and adopted as an additional feature for clustering. This answers to rationale (C).

By representing each phase difference $d\varphi$ with the value pair $[\cos(d\varphi), \sin(d\varphi)]$, we avoid discontinuities, at the cost of needing two clustering dimensions rather than one, for each phase difference. This answers to rationale (D).

For the clustering to work effectively, the numerical ranges of the values of all the clustered features need to be fairly similar. Therefore, the level differences are represented in Bels rather than decibels, relative to full scale, giving them a numerical range of about [-5…0]. This is a better match for the (cos, sin) values of the phase differences $\varphi_N - \varphi_1$, which are always in the interval [-1…1].

Finally, in order to enable the reconstruction of the estimated EGG waveform from the cluster centroid values, the phase φ_1 of the fundamental, too, is subjected to clustering. This answers to rationale (E). However, since the position of the cycle trigger point within the cycle is considered to be irrelevant, we do not want it to affect the outcome of the clustering. Therefore, the $\cos(\varphi_1)$ and $\sin(\varphi_1)$ values are down-weighted by a constant factor of 0.001, thus rendering negligible their contribution to the distances between the cluster centroids.

2.8 Statistical clustering

There are several clustering methods described in the literature. Here we are somewhat constrained by the demand for real-time operation and a continuously accumulating data set. Four clustering methods were compared [5]: K-Means, KMeansRT, Gaussian Mixture Model through expectation maximization and fuzzy C-means. In the end, a modified version was implemented of the KmeansRT algorithm [17]. It performs ‘hard’ clustering, i.e., each data point (corresponding to one EGG cycle) is assigned to one cluster only. Compared to other methods for clustering, the *k*-means method has these advantages: (1) it computes quickly, even in many dimensions, and (2) the number of points accumulated in the clusters affects only the centroid updates, not the classification. The latter means that, in a data set with thousands of EGG cycles, a small minority of cycles of an unusual shape can still give rise to a cluster of their own, especially if they occur early in the recording. A typical example is the weak sinusoidal cycles at onset and offset of voicing.

Two classical and as yet unsolved problems in statistical clustering are (1) how to determine the number of clusters that is in some sense optimal, and (2) how best to initialize or ‘seed’ the clusters for a relevant outcome. For the moment, we resolve these by trial-and-error, iterated until the clustering does not improve significantly with regard to the research question.

2.9 Sample entropy estimation

The ‘sample entropy’ of a signal is an interesting metric that has found numerous biomedical applications [18][19]. For brevity, it is often called SampEn. In studying phonation, we have found [3] that, with suitable parameter settings, the SampEn of cycle-synchronous data has the interesting property that it stays at effectively zero when phonation is stable, even with changing pitch, but peaks when ‘something out of the ordinary’ happens. This ‘something’ could be a voice break, such as those occurring between chest and falsetto voice, or other instabilities in phonation. In FonaDyn, the FDs form a vector of values that is updated with every new phonatory cycle. The number of FD magnitudes and FD phases that contribute to the SampEn metric is selectable, and typically smaller than for the waveshape analysis. The SampEn is computed over a running window of g glottal cycles, where the integer g can be chosen by the user, as can the sequence length l. Since our initial publication of the application of SampEn to cycle-based phonatory data [1], other research groups have also begun to use it in similar ways, e.g., [20]. The FonaDyn algorithm for estimating the SampEn is given in [5].

3 System description

3.1 Hardware requirements

The following hardware is necessary:

1. An electroglottograph device with skin surface electrodes and a signal conditioning box. There are about half a dozen models on the market.
2. A studio quality digital audio interface, accepting high level line input voltages, and with at least one mic preamplifier built in. (A laboratory data acquisition board will work only if it comes with audio drivers for the host computer, which is rarely the case. With FonaDyn, a frequency response down to 0 Hz is not needed, in which case audio interfaces perform to higher specifications than acquisition boards, at a lower cost.)
3. A passive voltage attenuator may be needed between the EGG device output and the audio interface input, to prevent clipping at the A/D converter.
4. A high-quality microphone, either head-mounted or stood at a fixed distance from the subject's mouth. The exact choice of microphone will depend on the ambient conditions and on whether or not the recordings are to be used for other analyses as well [23]. For FonaDyn, the choice is not critical, so long as a sufficient dynamic range is achieved.
5. A fast PC running Windows 7 or higher, Mac OSX 10.x or higher, or Linux (not yet tested).

Figure 2 shows a typical hardware setup, as a block diagram. This setup shows also some optional hardware that can be helpful if the operator and the subject are separated by a wall or window. If not, it can be simpler. A piano keyboard is convenient if prompting pitches are needed; or, these can be played from the computer. The audio interface can be external or internal to the computer.

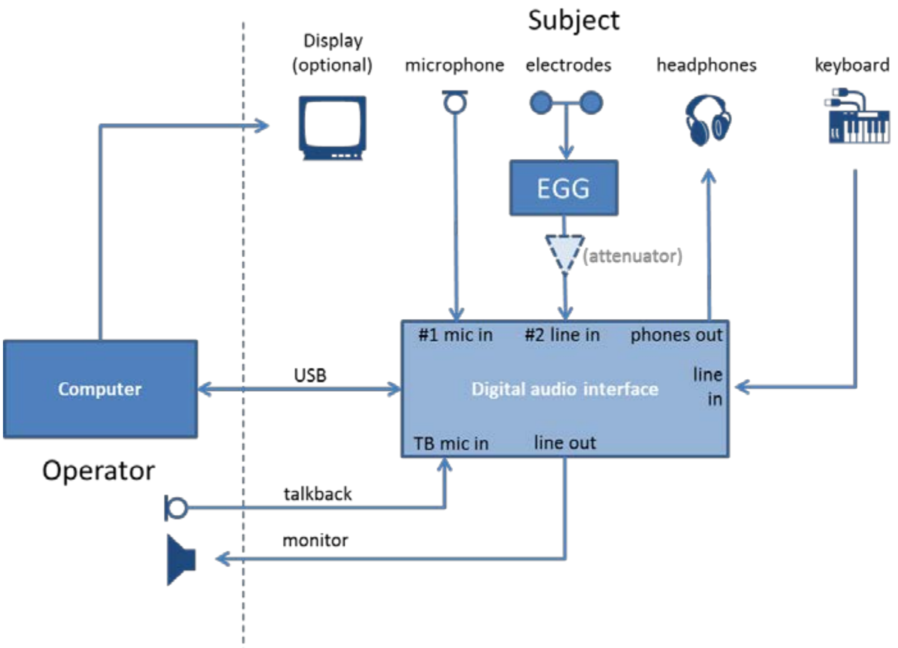


Figure 2. Typical hardware setup for FonaDyn.

3.2 Software architecture

FonaDyn is implemented in SuperCollider (SC) [21], a system for performing real-time sound analysis and synthesis. Originally developed by James McCartney, it is now maintained for and by a lively computer music community. Its users include also scientists, who have contributed a wealth of class libraries and plug-ins, making SuperCollider seriously useful also in an audio–music–acoustics research environment.

SuperCollider is open-source freeware that is supported on Windows, Mac OSX and Linux, and in reduced form on some other platforms as well. It has three major components: (1) a signal processing server, SCSYNTH, (2) an interpreted, object-oriented programming language SCLANG, acting as a client of the server, and (3) an integrated development environment, SCIDE, with an editor, control windows and help system. Code for the server and for user interaction and display are all written in SCLANG, which is similar to Smalltalk, with some idioms of other languages mixed in. SCLANG is profoundly object-oriented, and has many elegant, compact constructs for creative manipulation of arrays and collections, as applied in music composition. Although the SCLANG programmer hardly needs to know, the client and server communicate using Open Sound Control (OSC) messages, via a network protocol. For performance reasons, FonaDyn requires that the server and client processes be running on the same computer, but SuperCollider itself does not.

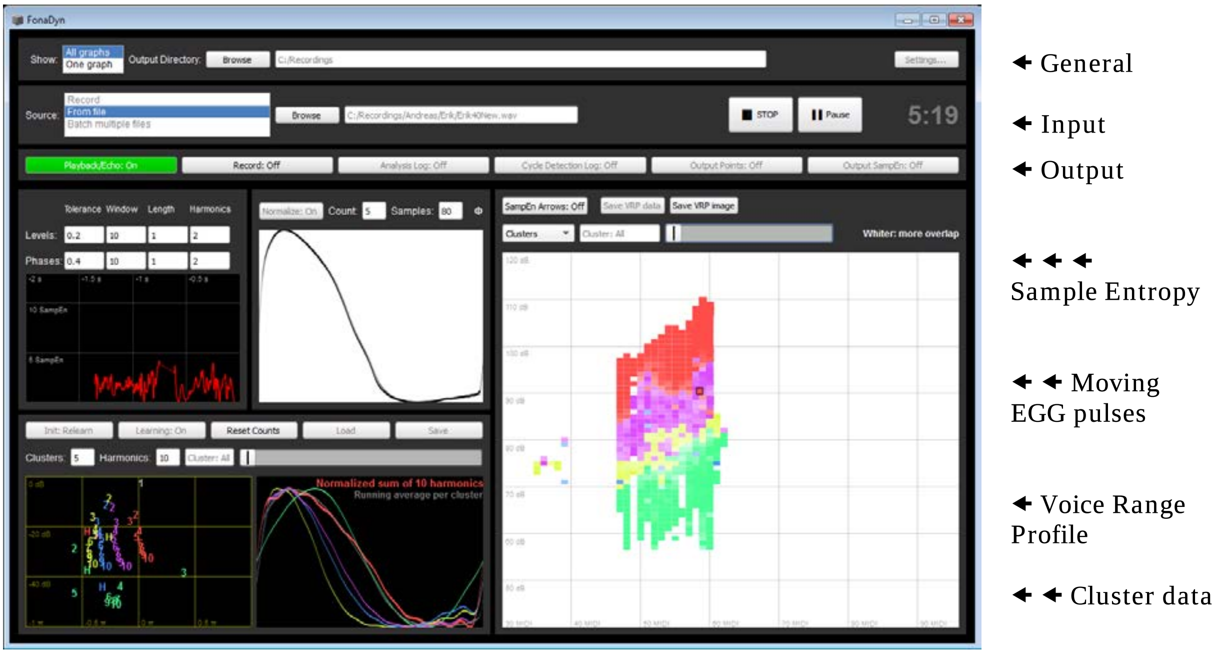
FonaDyn is not a stand-alone program, but a collection of classes and plug-ins that extend SCLANG. Hence FonaDyn is invoked from inside the SCIDE. The code structure of FonaDyn follows the model–view–controller paradigm, and is largely manifest in the layout of its main window (Figure 3). There are panels for General, Input, Output, Moving EGG, SampEn, Clusters, and VRP display. Each of these panels has a set of code class hierarchies related to its aspects of model, data, view, controller and signal processing.

When the user presses START, the schema for the signal processing is constructed dynamically on the server, according to the current settings, such as the numbers of harmonics, clusters and entropy parameters, and the input/output options. Increasing the number of harmonics will noticeably increase the computational load. Most of the graphics are redrawn continuously and completely at a fixed image rate of 24 Hz, whether the data has changed or not. New data is fetched from the server process at a rate of 60 ‘batches’ per second. The size of a batch will depend on how many glottal cycles have occurred.

The functions that control the rapid and gap-free exchange of control and data between the client and server processes are far from trivial, and took a major part of the development effort. FonaDyn now augments the standard interprocess communication mechanisms of SuperCollider. This is described in detail in [5], and in FonaDyn’s on-line Help files, which upon installation are integrated with the SuperCollider Help system.

3.3 User interface

The user interface is described in detail in the handbook that accompanies the FonaDyn download. Here we will give only an overview of the main features. The main window of FonaDyn (Figure 3) is rather like an instrument control panel, in that it has no pull-down menus, and most controls are visible all the time.

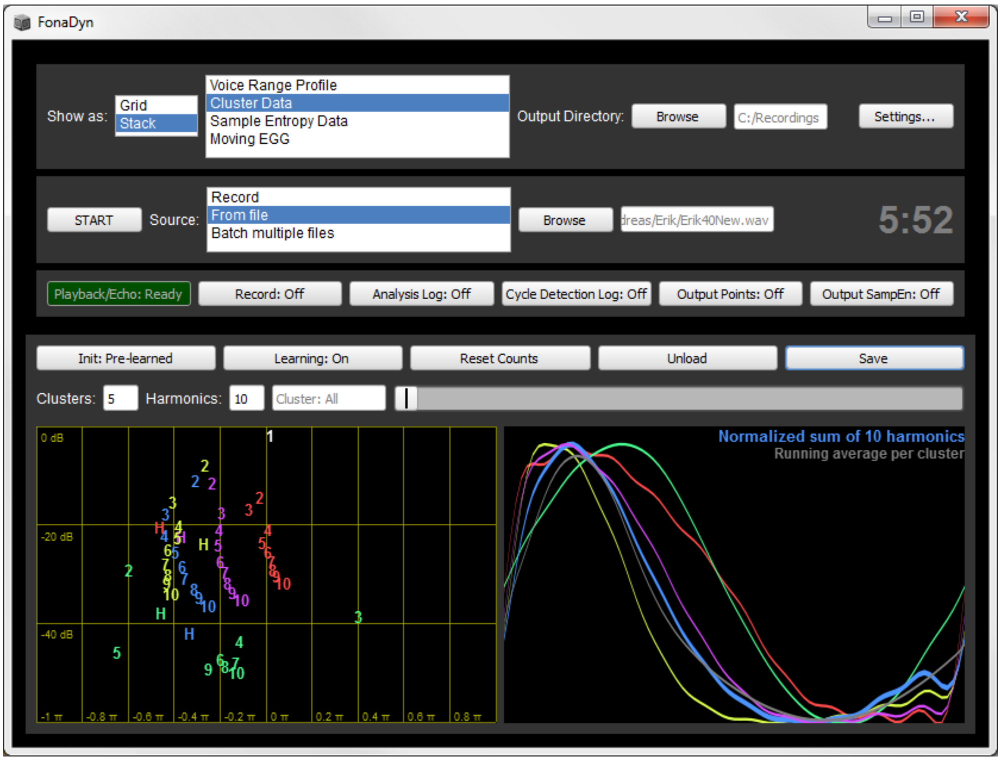


Show: All graphs
Output Directory: C:/Recordings
Source: From file
C:/Recordings/Andreas/Erik/Erik40New.wav
5:19
Playback/Echo: On
Record: Off
Analysis Log: Off
Cycle Detection Log: Off
Output Points: Off
Output SampEn: Off
Levels: Tolerance 0.2, Window 10, Length 1, Harmonics 2
Phases: Tolerance 0.4, Window 10, Length 1, Harmonics 2
Normalize: On
Count: 5
Samples: 80
SampEn Arrows: Off
Clusters: 5
Harmonics: 10
Cluster: All
Learning: On

Figure 3. The FonaDyn user interface in ’grid’ mode, showing all subpanels. All graphs change in real time.

The ‘Moving EGG’ display was inspired by the work of Christian Herbst [22], although our pulse segmentation works differently, and in real time.

By selecting the display mode ‘One graph’ rather than ‘All graphs’, the graphs are instead displayed one at a time. Display modes can be changed even while the program is running. For example, in Figure 4, the cluster data display panel has been selected.



Show as: Stack

Cluster Data

Output Directory: C:/Recordings

Source: From file

dreas/Erik/Erik40New.wav

5:52

Playback/Echo: Ready

Record: Off

Analysis Log: Off

Cycle Detection Log: Off

Output Points: Off

Output SampEn: Off

Init: Pre-learned

Learning: On

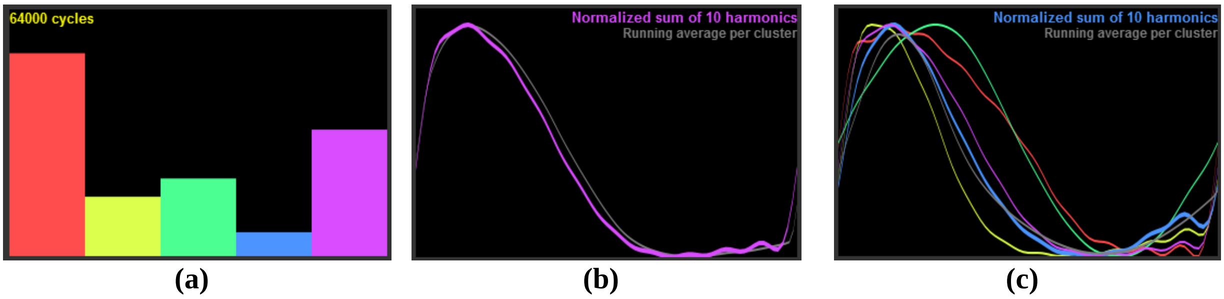
Clusters: 5

Harmonics: 10

Cluster: All

Figure 4. The main window in the more compact display mode 'Stack', with the Cluster Data panel selected (bottom half). At left, the colored centroids of the harmonics' relative levels ↓ and phases ↔. At right, EGG waveforms resynthesized from the centroids. Time → and amplitude ↑ are cycle-normalized; vocal fold contact area increases upwards. See also Figure 5.

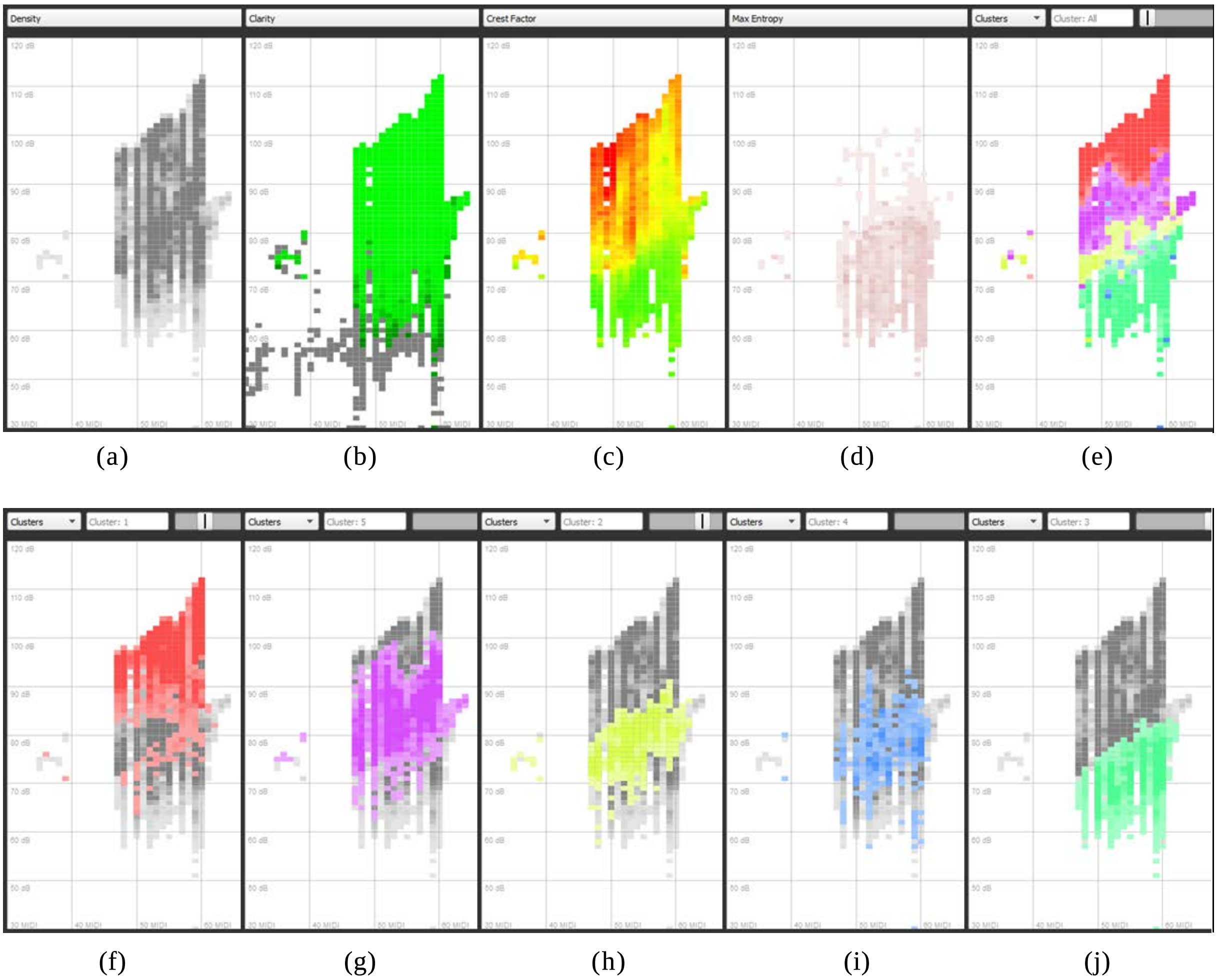
By clicking on the waveforms graph in the Clusters panel, the user can toggle between a bar graph, a single clustered waveform or all clustered waveforms (Figure 5).



(a) **(b)** **(c)**

Figure 5. Cluster statistics display. (a) Bar graph of EGG cycle counts, per cluster. (b) EGG wave resynthesis display, with one cluster selected. Gray line: a running average of recent pulses in the selected cluster. Purple line: EGG pulse resynthesized from the cluster centroids; with a close match to the average. (c) EGG wave resynthesis display, with all clusters selected. Both time and amplitude are cycle-normalized. Here, for example, the green signal (no vocal fold contact) is actually much weaker than the others.

The VRP panel can be switched to show one of several acoustic metrics, or the prevalence of each or all clusters. Figure 6 shows the different modes available, for the same data as in Figure 3. When 'learning', the association of cluster numbers to colors is arbitrary, and may change from on recording to the next, but when classifying, the color mapping stays the same.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 6. Partial screen dumps, to illustrate the modes of the VRP display. The horizontal axis is f_o , ten semitones/div; the vertical axis is sound level @ 0.3 m, 10 dB/div. The f_o , SPL, and metrics (a), (b) and (c) are derived from the audio signal, the rest from the EGG signal. A male amateur singer repeated soft-loud-soft /a/ vowels on several constant pitches over more than an octave. This recording took about 6 minutes; it is the same one as that in Figure 3. (a) 'Density', where darkest gray means ≥100 EGG cycles. (b) 'Clarity' showing accepted cycles (green) and rejected cycles (gray). (c) Crest factor (peak-to-RMS ratio) of the audio signal, where red means >12 dB. (d) Maximum SampEn; more brown means less stable phonation. (e) Dominant EGG waveshape cluster, by color; less saturation signifies more overlap between clusters. (f) - (j) Actual extent of individual cluster waveshapes (c.f. Figure 5c). The colors saturate at >50 EGG cycles in the cell. (f) Actual extent of 'red' cluster waveshapes; here, strongest phonation. (g) Actual extent of 'purple' cluster waveshapes; here, moderately strong phonation with full vocal fold contact. (h), (i) Actual extent of 'yellow' and 'blue' cluster waveshapes; here, soft phonation with less contact. (j) softest phonation (green) with no vocal fold contact, and a nearly sinusoidal EGG.

From Figure 6 it can be inferred, for instance, that the vocal folds can be vibrating without actually colliding even when the voice SPL is as high as 70-80 dB.

3.4 Input and output files

Input and output of audio and EGG signals were discussed in section 2.2. FonaDyn can write seven kinds of multi-track output files containing time-series data, including all the signals shown in Figure **1**. There is also text file output of cluster data and VRP results; and of bitmap image files (many formats) for saving VRP images. These files are intended to facilitate the further analysis and plotting of FonaDyn results, using spreadsheet apps or other math tools. Some example Matlab scripts for basic parsing and plotting of the output files are provided with the download. More detail is given in the accompanying handbook.

4 Application examples

4.1 Setup

Readers who wish to try FonaDyn are recommended first to download and read the first part of the handbook, where all aspects of setting up are described. Several software components need to be installed: the audio interface driver, the SuperCollider distribution matching the user’s computer; and the FonaDyn source code and precompiled plugins. No compilation is necessary, because SCLANG does that at run time. The only calibration needed is that of SPL. The location and arrangement of the hardware will depend on the study or purpose.

4.2 Mapping of voice registers

Our first paper on EGG waveform clustering [1] investigated the automatic discrimination between modal and falsetto voice, using Matlab. One example is given here, analyzed instead with FonaDyn. Figure 7 shows a VRP of three upward glissandi, each of about 6 s duration, by a male amateur subject. The figure shows the clustering obtained for 3 clusters analyzed with 10 FDs, on the second iteration over the 3×6=18 seconds. Here, blue corresponds closely to modal voice, red to falsetto, and green to a variant of falsetto which was audibly different from ‘red’. The subject breaks into falsetto at about 300 Hz, and then into the other variant of falsetto at about 500 Hz. The corresponding EGG waveshapes are also shown. Figure 8 shows Matlab plots of a FonaDyn Log File output, from the same run of three glissandi.

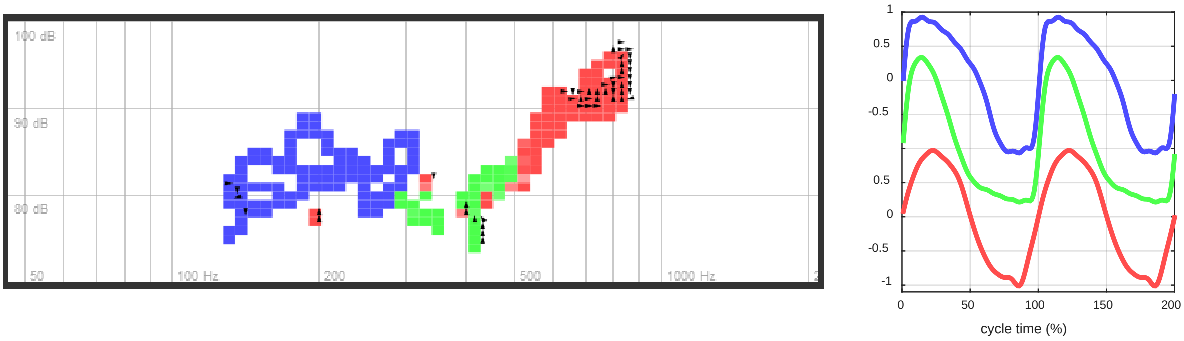


Figure 7: Voice register mapping. Left: VRP of 3 upward glissandi by a male amateur singer, with 3 clusters based on 10 FDs; clusters resulting from two iterations over the 18 s. Blue – modal voice, red and green – variants of falsetto. The rectangular cells are one semitone wide and one dB high. Black pointers indicate where the SampEn was high. The view is cropped, for clarity. Right: resynthesis of the corresponding EGG waveshapes, from 10 FD centroids. EGG pulses are cycle normalized in both time and amplitude. Modal voice has the longest closed phase, green the shortest.

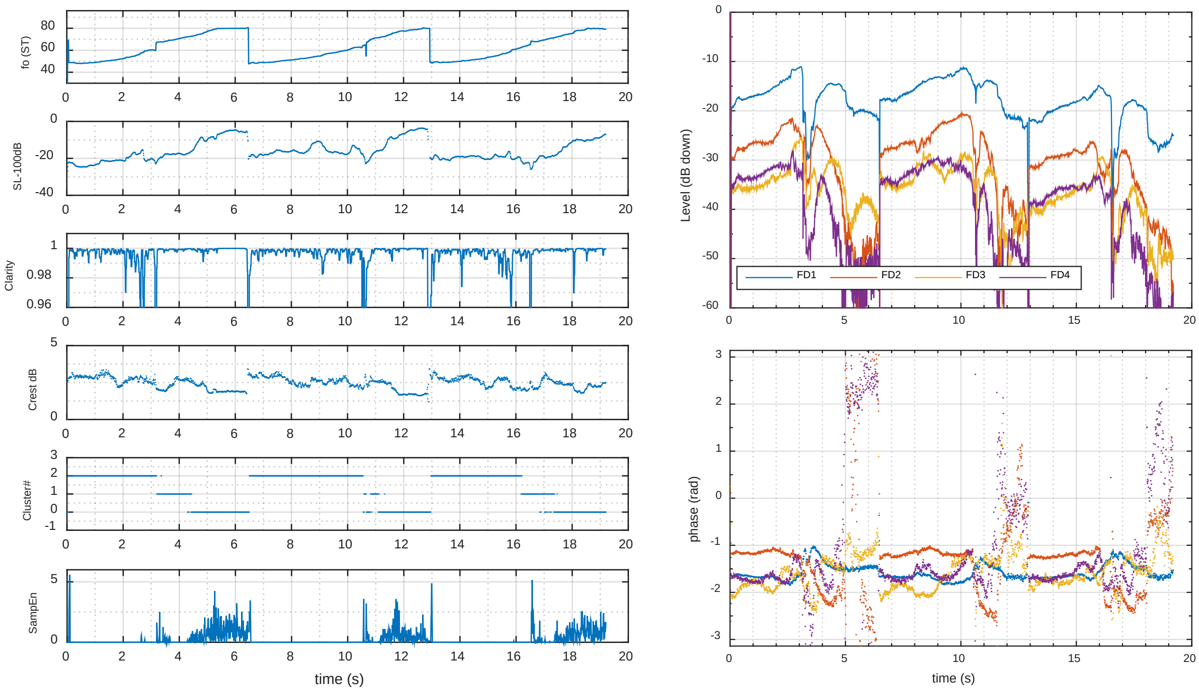


Figure 8. Log file output of analysis results, 100 Hz frame rate; plotted in Matlab and scaled for the preceding example. Left from top: f_o showing three upward glissandi, with voice breaks near 3.2, 10.7 and 16.5 s; sound level; 'clarity' metric; crest factor; cluster number, here 0=red, 1=green and 2=blue (colors of Figure 7); and the SampEn metric. Right: level and phase plots of the first 4 (of 10) Fourier Descriptors, corresponding to harmonics of the EGG signal.

4.3 Giving feedback on voice use

In the singing studio, it can sometimes help to visualize a particular aspect of voice production. One such aspect is pressedness, that is, the amount of adductive force with which the vocal folds are held together; in most genres, excessive press is undesirable and possibly risky. A first study on this matter [6] investigated whether it would be possible to provide feedback to the singing student of the degree of pressedness. Four trained female singers were asked to phonate both normally and with excessive press, as monitored by a teacher, on low and high pitches, while remaining in chest voice. Figure 9 shows examples of typical normal and pressed cycle shapes for high-pitched and low-pitched productions. It should be noted that the ground truth here was the singer's *intent* to use normal or pressed phonation, so there remains some uncertainty regarding how these modes were defined, and executed.

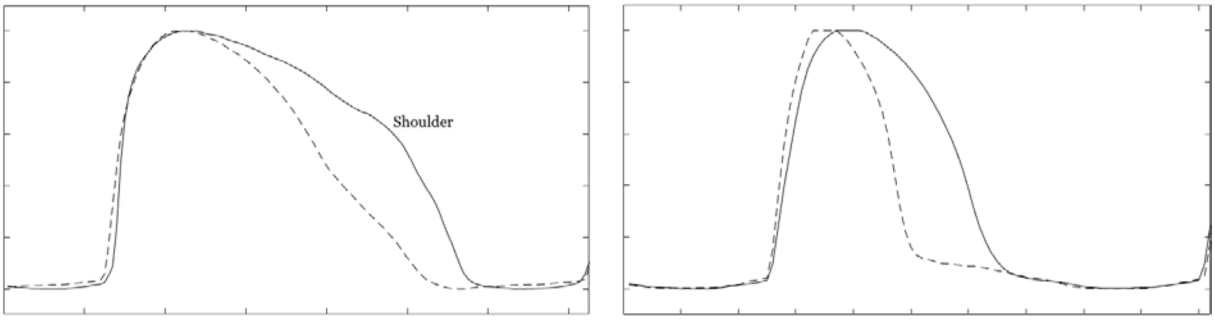


Figure 9. Example EGG waveforms of normal (dashed) and pressed (solid) phonation; female trained singer. Left: comfortable low pitch, right: comfortable high pitch. From [6], used with permission.

For both high- and low-pitched files, pressed quality cycles were wider, corresponding to a longer closed phase of the cycle, and more asymmetric due to increasing differences between closing and opening phases. In general, it was not possible to discriminate pressed from normal automatically using only two clusters, however, with a strategy of several clusters that were whittled down manually to two, correct intra-subject classification rates of around 90% (of EGG cycles) were achieved, such that the on-screen visual feedback felt fairly reliable. Inter-subject classification gave mixed results, as might be expected.

4.4 Assessing effects of therapy or training

In Section 3.3, we used example data from [4] for visualizing how the EGG varies over the voice range. The clustering obtained in a pre-treatment session can be retained, and used to classify the production of the same subject, post treatment. This can show interesting aspects of how the phonation has changed. Figure 10 gives such an example, from Lã [24]. A male singer produced a baseline VRP, by making soft-loud-soft vocalizations on constant pitches, then did a particular exercise, and then repeated the VRP. The VRP post treatment shows higher boundaries of blue, and lower of green. With experience of EGG waveshapes, this can be taken to mean that, after the exercise, the subject sang louder without going ‘pressed’, and softer without going ‘breathy’. Whether or not these changes were actually due to the exercise is of course a different question; as is determining the statistical significance of the differences. Note that the color maps here are created from data from several tens of thousands of EGG cycles, as phonated during the two or three minutes of the VRP recording, so the pre-post change that is manifest in this particular trial is beyond question.

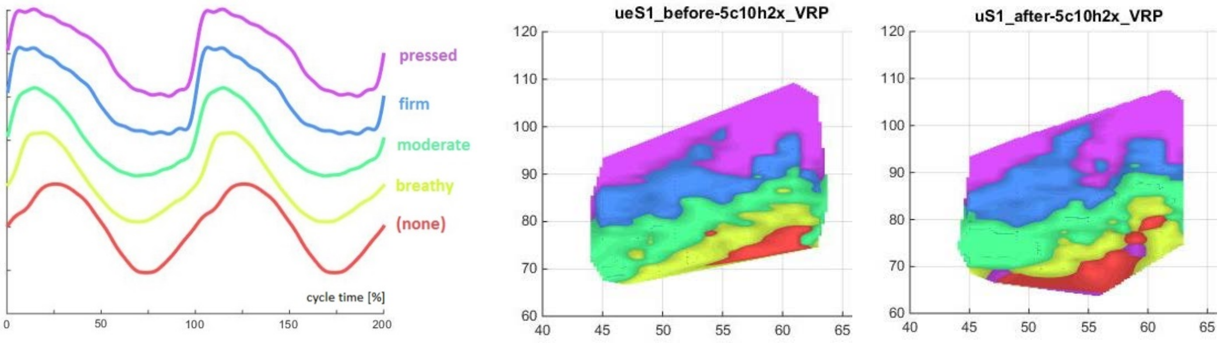


Figure 10. Example of visualizing an effect of a treatment. Left: pre-treatment EGG waveshapes, cycle-normalized, interpreted as varying degrees of vocal fold adduction. Middle: pre-treatment, accumulated distribution over pitch (horizontal, in semitones) and SPL (vertical, in dB @ 0.3 m). Right: post-treatment distribution of the same waveshapes. The outcome is discussed in the text. Data from [24], rendered here with permission. These graphs were customized in Matlab, using CSV data files from FonaDyn.

5 Status report

5.1 Operating system support

FonaDyn runs on Windows from version 7 and on OSX from version 10.x. The analysis results are identical on these two platforms. SuperCollider is supported also on Linux. At this writing, a Linux version of FonaDyn has been compiled, but not tested.

While SCLANG is platform-independent, SCSYNTH is not; in particular, plugins including those particular to FonaDyn must be compiled specifically for each platform. On Windows, the implementation is 32-bit, but it will run without problems on 64-bit Windows. We intend eventually to compile a 64-bit version of FonaDyn for testing, but will not release it separately unless it performs significantly better.

5.2 Strengths and weaknesses

FonaDyn’s strength are that it offers a new and rich view on the EGG. It is freeware, as is the supporting code, and it can be tweaked by the SC-literate user. It has a rich variety of outputs for use by researchers, and its real-time operation will be of interest to both voice clinicians and voice pedagogues.

The currently released version 1.3.5 has the following known problems. In normal use, these limitations are not a large concern.

1) On long runs of more than, say, 7-10 minutes, operation may become sluggish and eventually stop. We believe this to be an intrinsic problem of the heap management, or ‘garbage collection’, in SuperCollider, and hope to find a solution soon. When a run has completed, it can take several seconds for the program to return to the ground state, presumably while the internal ‘grey-list manager’ is tidying up the memory heap. The faster the PC, the smaller this problem becomes.
2) If the user interface is manipulated too briskly during operation, the client-server handshaking may stumble, causing the program to stop.

5.3 Performance

Some examples of CPU load figures are given in Table 1, for a typical case of 5 clusters and 10 harmonics. The client load decreases somewhat if not all graphs are displayed. The loads are those reported by the Windows Task Manager. They will vary somewhat also with the choice of audio interface, depending on its driver software. The figures below were obtained with an external audio interface (RME Fireface UCX via USB). Since some of the client-server data exchange has to be done through many small temporary files, a solid-state disk drive is recommended.

Table 1. Typical CPU loads on Windows 7; for 5 clusters and 10 harmonics.

Computer	OS	CPU load SCSYNTH (server)	CPU load SCLANG (client)	Usable time for one recording
Core 2, 2.8 GHz PC tower, hard disk	Win7 32-bit	4-6%	18-27%	6-7 minutes
i7, 3.0 GHz laptop, SSD	Win7 64-bit	1-2%	7-15%	8-12 minutes

FonaDyn does the full processing as described above, for every individual EGG cycle. At very high f_o , the number of cycles per second can become difficult for the program to handle. This impacts mostly the updating of the display, which can become jerky; we have not seen data being lost. If the client’s event scheduler queue becomes too long, a warning is printed and the program stops benignly. A future optimization might be to skip some cycles at high f_o , since those cycles usually are very similar to each other.

5.4 Lessons learned

EGG waveform clustering can be used not only to classify automatically different types of phonatory contact patterns, but also to stratify automatically continua in EGG pulse shape. This introduces a new paradigm in voice analysis.

The number of clusters, and their initialization, are very important, and must be attuned to each specific research question or application. In most cases, the pragmatic solution will be to record a representative set of productions first, and to analyze those recordings iteratively, until a clustering that fits the question is found. This clustering can then be used to classify new recordings. A case example of this procedure is described in detail in [6]. For each research question, a specific method for subsequently assessing the statistical significance of differences between VRPs will be needed.

5.5 Mode of availability

FonaDyn is available in source code form under EUPL (European Union Public Licence) v1.2, by an e-mail request for a download link, to the first author. Any and all use of FonaDyn and derivatives thereof must be fully acknowledged.

5.6 Future plans

It would be interesting to replace the EGG signal with some other periodic signal derived from phonation, for instance, the signal from an accelerometer placed on the neck, or a photoglottographic signal (PGG), and then use FonaDyn in exactly the same way, or indeed with several simultaneous signals. The glottal area (as from the PGG) is complementary to the EGG, in that it provides most of its information in the open phase of the glottal cycle, rather than in the closed phase. A study on this is under way at this writing.

We intend also to explore further the clustering paradigm more generally in vocology, and to use FonaDyn for continued research on analysis of voice production, with clinical and pedagogical applications. Comments and suggestions for new features are received gratefully, but without commitment. There are no plans for commercial exploitation.

Acknowledgments

FonaDyn stems from work funded by the Swedish Research Council, projects 2010-4565 and 2013-5194, as well as by KTH faculty grants. Most of the principles underlying FonaDyn were developed by Andreas Selamtzis, in the course of his doctoral studies. The foundation version of the real-time SuperCollider implementation of FonaDyn was written by Dennis Johansson, for his M.Sc. degree project in computer science. Isak Nilsson made important contributions to the OSX implementation. The continued development has been done by Sten Ternström, who was also the supervisor of Johansson and Selamtzis. We are indebted to many of our colleagues, especially Peter Pabon, Christian Herbst and Matthias Echternach, for valuable inspiration and discussions. We express our thanks to the volunteering subjects and voice professionals who participated in the various pilot studies, especially to Ragnhild Bøhn Sørby, Grethe Kristine Grindhaug Brekke, Katri A. Keskinen and Jeanette LoVetri.

References

[1] Childers, D., Larar, J.N. (1984). Electroglottography for laryngeal function assessment and speech analysis. *IEEE Trans. Biomed. Eng.*, BME-31 (12), Dec 1984, 807-817.

[2] Selamtzis, A. (2014). *Electroglottographic analysis of phonatory dynamics and states*. Licentiate thesis in Speech and Music Communication, Stockholm: KTH Royal Institute of Technology, 2014. , vii, 31 p. Available online at http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-145692

[3] Selamtzis, A., and Ternström, S. (2014). Analysis of vibratory states in phonation using spectral features of the electroglottographic signal. *J. Acoust. Soc. Am.*, 136(5), 2773-2783.

[4] Selamtzis, A., and Ternström, S. (2016). Investigation of the relationship between the electroglottogram waveform, fundamental frequency and sound pressure level using clustering. *J. Voice*, available online at http://dx.doi.org/10.1016/j.jvoice.2016.11.003.

[5] Johansson, D. (2015). *Real-time analysis, in SuperCollider, of spectral features of electroglottographic signals*. M.Sc. degree thesis in computer science, KTH Royal Institute of Technology, Stockholm, Sweden. Available online at this link (October 2016).

[6] Nilsson, I. (2016). *Electroglottography in real-time feedback for healthy singing*. M.Sc. degree thesis in computer science and communication, KTH Royal Institute of Technology, Stockholm, Sweden. Available online at this link (December 2016).

[7] Ternström, S., Pabon, P., and Södersten, M. (2016). The Voice Range Profile: its function, applications, pitfalls and potential. *Acta Acustica united with Acustica*, 102(2), 268–283.

[8] Roubeau, B., Henrich, N., and Castellengo, M. (2009). Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited. *J. Voice*, 23 (4), July 2009, 425–438.

[9] *Matlab* © The MathWorks, Inc. www.mathworks.com

[10] Herbst, C. and Ternström, S. (2006) A comparison of different methods to measure the EGG contact quotient. *Log. Phoniatr. Vocol.*, (31) 126-138. DOI: 10.1080/14015430500376580.

[11] Mooshammer, C. (2010) Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German. *J. Acoust. Soc. Am.* 127:1047–1058.

[12] Herbst, C.T., Fitch, W.T.S, Švec, J.G. (2010). Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasively. *J. Acoust. Soc. Am.*, 128 (5), 3070-3078.

[13] Titze, I.R. (1989). A four-parameter model of the glottis and vocal fold contact area. *Speech Communication*, 8 (3), 191-201.

[14] Titze, I.R. (1990). Interpretation of the electroglottographic signal. *J. Voice*, 4 (1), 1-9.

[15] Dolanský, L.O. (1955). An Instantaneous Pitch-Period Indicator. *J. Acoust. Soc. Am.* 27, 67-72 (1955); http://dx.doi.org/10.1121/1.1907499

[16] McLeod, P. and Wyvill, G. (2005). A Smarter Way to Find Pitch. *Proc Int'l Computer Music Conf*; ICMC 2005, 138-141. Permalink: http://hdl.handle.net/2027/spo.bbp2372.2005.107 . [An

implementation of the above, called “Tartini”, is included with the ‘SC3-plugins’ library of signal function blocks]

[17] McFee, B (2012). *More like this: machine learning approaches to music similarity*. PhD thesis, University of California at San Diego, 186 p (algorithm B.1, p. 152). Available online at http://bmcfee.github.io/papers/bmcfee_dissertation.pdf. [An implementation of the above, called “KMeansRT”, is included with the ‘SC3-plugins’ library of signal function blocks. FonaDyn supplements this with “KmeansRT2”, which provides the option of continuing learning with a pre-learned vector of centroids.]

[18] Richman, J.S., Randall Moorman, J. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology*, 278 (6), 2039-2049.

[19] Yu-Hsiang Pan, Yung-Hung Wang, Sheng-Fu Liang, Kuo-Tien Lee (2011). Fast computation of sample entropy and approximate entropy in biomedicine. *Computer Methods and Programs in Biomedicine*, 104 (3), 382-396.

[20] Echternach M, Burk F, Köberlein M, Burdumy M, Döllinger M, Richter B. The influence of vowels on vocal fold dynamics in the tenor’s passaggio. *J. Voice*, in press. http://dx.doi.org/10.1016/j.jvoice.2016.11.010

[21] SuperCollider website: http://supercollider.github.io/

[22] Herbst C, 2004. MovingEGG. Available online at this link. (Accessed 14 Oct 2015).

[23] Švec JG, Granqvist S (2010). Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*, 19, 356–368, November 2010.

[24] Lã FBM. Short-term effect of an aerodynamic exercise on phonation in singing, as assessed by the EGG waveform. Personal communication, 2016.

This table of contents is included only for the convenience of the reviewers.

Contents