KTH Computer Science
and Communication

Electroglottographic analysis of phonatory dynamics and states

ANDREAS SELAMTZIS

Licentiate Thesis
Stockholm, Sweden 2014

TRITA-CSC-A 2014:09
ISSN 1653-5723 KTH School of Computer Science and Communication
ISRN KTH/CSC/A-14/09-SE SE-100 44 Stockholm
ISBN 978-91-7595-189-8 SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framläggs till offentlig granskning för avläggande av teknologie licentiatexamen i tal- och musikkommunikation, fredagen den 13 juni 2014 klockan 15.15 i sal Fantum, Lindstedtsvägen 24 plan 5, Kungl Tekniska högskolan, Stockholm.

© Andreas Selamtzis, juni 2014

Tryck: Universitetsservice US AB

Abstract

The human voice is a product of an intricate biophysical system. The complexity of this system enables a rich variety of possible sounds, but at the same time poses great challenges for quantitative voice analysis. For example, the vocal folds can vibrate in several different ways, leading to variations in the acoustic output. Because the vocal folds are relatively inaccessible, such variations are often difficult to account for. This work proposes a novel method for extracting non-invasively information on the vibratory state of the human vocal folds. Such information is important for creating a more complete voice analysis scheme. Invasive methods are undesirable because they often disturb the subjects and/or the studied phenomena, and they are also impractical in terms of accessibility and cost. A useful frame of reference for voice analysis is the Voice Range Profile (VRP). The 3 dimensional form of the VRP can be used to depict any phonatory metric over the 2 dimensional plane defined by the fundamental frequency of phonation (x-axis) and the sound pressure level (y-axis). The primary goal of this work was to incorporate information on the vibratory state of the vocal folds into the Voice Range Profile (e.g., as a color change). For this purpose, a novel method of analysis of the electroglottogram (EGG) was developed, using techniques from machine learning (clustering) and nonlinear time series analysis (sample entropy estimation). The analysis makes no prior assumptions on the nature of the EGG signal and does not rely on its absolute amplitude or frequency. Unlike time-domain methods, which typically define thresholds for quantifying EGG cycle metrics, the proposed method uses information from the entire cycle of each period. The analysis was applied in a variety of experimental conditions (constant vowel with different vibratory states, constant vibratory state and different vowels, constant vowel and vibratory state with varying lung volume) and the magnitude of effect on the EGG short-term spectrum was estimated for each of these conditions. It was found that the short-term spectrum of the EGG signal sufficed to discriminate between different phonatory configurations, such as modal and falsetto voice. It was found also that even supposedly purely articulatory changes could be traced in the spectrum of the EGG signal. Finally, possible pedagogical and clinical applications of the method are discussed.

Acknowledgments

A report like this is the perfect opportunity to thank all the people that contributed to this work or to my well being. In strict non-alphabetical order I would like to thank from the bottom of my heart:

My supervisor Sten Ternström, for providing me large quantities of inspiration, motivation and support. My co-supervisor Olov Engwall for promptly giving his detailed and helpful feedback. The head of the TMH department, Anders Askenfelt who was always willing to answer scientific questions on music acoustics, and for entrusting me with the key to the grand piano that is owned by TMH and sits in the Operahögskolan. Johan Sundberg, the grandfather of singing voice research, who is always available and willing to share his expertise. Peter Pabon, major contributor of the FonaDyn project, who was very helpful by providing ideas, input and technical help whenever needed. Christian Herbst for openly sharing his ideas and giving important input on EGG processing. Special thanks go to the test subjects who willingly took part in the experiments.

Gaël Dubus and Anders Elowsson, my office mates, for being so patient with me vocalizing pitch glides in the office, while making trial EGG recordings. My office neighbor, Anders Friberg who is always there to discuss research and music. Giampiero Salvi, for generously providing valuable input on statistical modeling and clustering. Roberto Bresin for stimulating conversations on research, music performance and culture. Glaucia Salomão for her friendly and outward temperament and all the discussions we had on life and science. Marc Arnela Coll, for giving his very helpful feedback on my presentations and sharing his thoughts and feelings on doing research. Chris Koniaris for his friendship and his help with both scientific and practical matters on living in Sweden.

Fellow students and colleagues at TMH from the SMC and Speech group, Luca, Laura, Jana, Catha, Raveesh, Kalin, Martin, Niklas, Simon, Bajibabu, Jens, Samer, Meg, Kjetil, Ze, Sofia, David and Johan for making this building a positive environment to work at. Anita Kruckenberg for being such a positive, enthusiastic character and for helping me in learning Swedish.

My lifelong friend Ioanna for standing by my side during critical times. My family for providing the foundations of my existence material-, intellectual- and spiritual-wise.

This work has been carried out with the support of the FonaDyn project, Vetenskapsrådet contract number 2010-4565.

Contents

Acknowledgments **v**

Contents **vii**

1 Introduction **1**
1.1 Basics of phonatory function 1
1.2 Experimental methods in voice science 2
1.3 Non-invasive voice analysis 3
1.4 The Voice Range Profile 4
1.5 Phonatory dynamics and states 5
1.6 Problem statement 9

2 Methods **11**
2.1 Analysis of the electroglottogram 11
2.2 Entropy measures in voice analysis 14
2.3 Clustering using the Relative Fourier Descriptors 15
2.4 Outline of the proposed method 15

3 Results **19**
3.1 Summary of results of the included paper 19
3.2 Sources of variation in the EGG 19

4 Discussion **27**

Bibliography **29**

Paper I **33**

1 Introduction

1.1 Basics of phonatory function

The voice organ is composed of the lungs, the trachea, the larynx and the vocal tract. A short description of each of these parts is given below:

- The **lungs** are organs made of sponge-like tissue that can store air,
- The **trachea** is the airway tube that connects the lungs to the larynx,
- The **larynx** or voice box is a construction of muscles and cartilages that controls the various degrees of freedom of voicing. The main sound producing component of the larynx are the **vocal folds**: two stripes of soft tissue, each positioned on either side of the median plane[1].
- The **vocal tract** is the space between the vocal folds and the lips. The most important property of the vocal tract for voicing, is its deformability by means of moving articulators (tongue, jaw, e.tc.).

The space between the two vocal folds is called the glottis. When the vocal folds are separated (abducted), the glottis is fully open and the air is transported freely into and out of the lungs, as in inhalation and exhalation. When specific laryngeal muscles are activated, the vocal folds come closer together (adduct) and the glottis becomes narrower. Under this condition, the passing airflow can drive the vocal folds into vibration. The most common way of vibration of the vocal folds is a periodic motion towards and away from each other. When the vocal folds approximate, the area of the glottis decreases until the vocal folds collide, and then increases until the vocal folds reach their maximum excursion from their prephonatory position. The effect of this vibration on the airflow is the generation of acoustic waves, which propagate through the vocal tract, and ultimately radiate from the mouth.

Although voicing occurs spontaneously ever since the moment of birth, speech, like every skill, requires a process of training. The complex structure of the biophysical voice organ calls for sophisticated neuromuscular control. By simultaneously

[1] The median plane is the plane that separates our body in two externally symmetric left and right halves.

controlling the laryngeal muscle tensions, the lung pressure and the shape of the vocal tract by means of the articulators, humans are able to speak and sing, communicating meanings and emotions.

1.2 Experimental methods in voice science

Our understanding of the physical mechanisms of human phonation has advanced greatly during the past century. Several experimental techniques exist that provide insights into voice function: electromyography, magnetic resonance imaging, high-speed videoendoscopy, photoglottography, neck surface acceleration measurements, acoustic analysis, electroglottography. These techniques can be briefly described as follows:

- **Electromyography** can be used to measure activation patterns of the respiratory and laryngeal muscles that control phonation.
- **Magnetic resonance imaging** is used to record in ever-increasing detail the 3D geometry of the vocal tract.
- **High-speed videoendoscopy** is applied in filming the vibration of the vocal folds as seen from above by insertion of an endoscope through the mouth or the nose.
- **Photoglottography** monitors the amount of light transmitted through the glottis by placing a light source either below or above the vocal folds and a light sensor on the other side.
- **Acoustic analysis** and signal processing is done on the airborne voice signal which is acquired by one or several microphones at a distance from the subject's mouth.
- **Electroglottography** is used to monitor the electrical impedance across the neck by means of two electrodes placed externally at the level of the larynx. The impedance increases slightly when the vocal folds are separated and decreases when they are in contact.
- **Neck surface acceleration** is measured externally by placing an accelerometer at a point that is close to the glottis (near the thyroid cartilage) or above the collarbone.

Still, there remain many open questions in voice research, as today's technology does not allow an *in-vivo* quantification of all the neuromuscular and mechanical phenomena that take part in vocalization. Of the above techniques, the first four can be used only in a clinical environment, and with the exception of MRI, they are invasive. The advantage of the invasive techniques is that they provide detailed

data that reflect the behavior of specific parts of the voice organ. The disadvantage is that the presence of equipment may cause discomfort to the test subject, and may restrict phonation. For example, although high-speed videoendoscopic examinations are very effective in a clinical context, their potential is reduced when studying the singing voice; an endoscope in the oral cavity does not allow the subject to vocalize freely, as it obstructs a large part of the vocal tract. A less invasive way to obtain an image of the vocal folds as they vibrate is a fiberscope inserted through the nose. A limitation of the fiberscope though is its reduced pixel resolution compared to the rigid endoscope.

1.3 Non-invasive voice analysis

If the experimental protocol requires the test subject to be able to vocalize freely non-invasiveness is a requirement. The least invasive techniques that can be utilized are acoustic analysis, electroglottography (EGG) and measurement of neck surface acceleration.

The product of the voice organ is the acoustic signal that is radiated from the mouth opening. Being the final element in the chain of voice production, it is affected by the behavior of all the preceding parts. The acoustic signal can be modeled in terms of the source-filter theory (Fant, 1960). In short, the source refers to the transglottal airflow, i.e., the acoustic volume velocity at the exit of the vocal folds and inlet of the vocal tract. The filter refers to the resonant vocal tract, which acts as an acoustic filter to the source. When the vocal folds vibrate periodically, the spectrum of the voice source signal is harmonically structured and the strength of the harmonics decreases monotonically with frequency. The vocal tract being a resonant cavity, provides peaks (formants) and valleys (zeros) in specific frequencies that modify the amplitudes of different harmonics of the source spectrum. The location of the peaks and valleys along the frequency axis is adjusted by the arrangement of the vocal tract. Under certain conditions, the acoustic signal can be processed to recover the characteristics of the source by counteracting the effect of the vocal tract. This procedure is called inverse filtering; however to the author's knowledge no robust automated algorithm for inverse filtering exists that will handle the entire voice range in F0 and SPL.

Electroglottography provides a relative measure of the contact area between the two vocal folds. For a detailed review on electroglottography see (Baken, 1992). The measurement principle exploits the electrical conductivity of the vocal folds and the tissue and cartilages that surround them. Two electrodes are placed on the neck surface and a high-frequency current is passed between the electrodes. When the vocal folds vibrate, they come in contact and the current that passes through them increases, as the electrical impedance across them is decreased. Conversely, when the vocal folds separate during the vibration cycle, the current that passes through them diminishes. Thus the resulting waveform represents the contact area of the vocal folds from a minimum to a maximum value.

As the acoustic signal excites the tissue surrounding the larynx, an accelerometer externally placed on the surface of the neck can pick up the vibrations. The accelerometer signal retains some properties of the original exciting acoustic signal, however it is filtered and attenuated by the intermediate tissue. It is a useful signal for monitoring voice use, as it is convenient to acquire and extract from it metrics that correlate to those extracted from the airborne signal (Mehta et al., 2012).

1.4 The Voice Range Profile

Two prominent features that can be extracted from the acoustic signal are the sound pressure level (SPL) and the fundamental frequency of phonation (F0). The perceptual cues that closely link to these features are loudness and pitch. It should be noted however loudness apart from being correlated to SPL, varies with the spectral content of a given sound. The intention of a subject to phonate at a specific fundamental frequency with a specific amount of effort is translated by the nervous system into activation of the different parts of the phonatory system which finally gives rise to the desired phonation.

A *Voice Range Profile* (VRP) or phonetogram is a measurement method where a test subject exercises the range of fundamental frequencies and sound pressure levels that are possible to produce. The result can be plotted on a two dimensional graph, with the fundamental frequency F0 on the x-axis (logarithmic scale) and the SPL on the y-axis. A third dimension can be used to map another parameter such as the crest factor (ratio of peak over rms amplitude) of the acoustic waveform, jitter, or frequency of occurence for a given (F0, SPL) point. In regular phonation the estimates of F0 and SPL can be taken pitch-synchronously, meaning that for each fundamental period of the voice signal there will be an (F0, SPL) point. If the EGG signal is recorded simultaneously, parameters of the EGG waveform can also be plotted on the VRP plane.

When acquiring a VRP, the voice of the subject might vary considerably in different (F0, SPL) regions, in terms of timbre or quality, or in terms of the subject's self-sensory feedback on their own use of the voice. Such changes can emerge from differences in the vibratory state of the vocal folds or adjustments in the supraglottal tract. These changes introduce a source of variation that needs to be taken into account. An example of a source of variation in the VRP might be the modal and falsetto registers. Typically in an upward pitch glide of an untrained subject, the voice will switch at a given (F0, SPL) point from chest voice to falsetto voice. Conversely in a downward pitch glide, there will be a switch from falsetto to modal, though not at the same (F0, SPL) point, due to hysteresis (Švec et al., 1999). Apart from modal and falsetto, several other vibratory states exist. Accounting for them would be of central importance for an improved interpretation of the VRP.

1.5 Phonatory dynamics and states

1.5.1 Nonlinearity

An essential element of the vocal folds as a mechanical system is its nonlinearity (Davis and Fletcher, 1996). An introduction to nonlinear dynamical systems can be found in (Strogatz, 2006). Contrary to linear oscillators, nonlinear oscillators exhibit a dynamical behavior depending on their internal state as well as the external forcing applied to them. In nonlinear oscillating systems, the external forcing is a function of the motion variables (e.g., position and velocity) of the oscillator. This relationship between the forcing and the motion of the oscillator is the cause of self-sustained oscillation: an initial change in the position and velocity of the system starts the oscillation, which is maintained by the dynamically adapted forcing.

As a nonlinear dynamical system, the vocal folds can vibrate in many different ways. A readily identifiable way of vibration, common in speech and many singing styles is stable periodic motion, when the vocal folds oscillate with a specific fundamental frequency and small pertubations around that frequency. However, many other vibratory phenomena are possible during phonation. For example, a voice during singing or emotionally stressed speaking may "break" as the fundamental frequency of phonation unintentionally jumps to a higher value. Such abrupt changes can be associated to bifurcations (Lucero, 1996; Švec et al., 1999), which appear only in nonlinear dynamical systems.

Other nonlinear phenomena that are encountered in phonation include period-multiplying and deterministic chaos (Neubauer et al., 2004; Herzel et al., 1995; Jiang et al., 2006). Period multiplying in phonation refers to the sudden change in the vibratory period of the vocal folds such that the new period of vibration is an integer multiple of the preceding period. Deterministic chaos is a type of motion that looks random despite the underlying order that guides its evolution. While random systems are truly impossible to predict, the evolution of a chaotic system can be predicted, though for a limited time span, depending on the accuracy with which its initial conditions are known.

Phenomena of hysteresis are also a result of nonlinearity. For example, while initiation of vocal fold vibration may require a given amount of subglottal pressure P_{sub}, once the oscillation starts, the subglottal pressure typically can descend to lower values without the oscillation having to stop (Plant et al., 2004; Lucero, 2005). Hysteresis may play a crucial role in professional vocal control, e.g., when a singer negotiates a register break, as it implies that regulation of phonatory control depends on the history or time record of a phonatory production.

1.5.2 Phonatory dynamics

Phonation requires the coordination of different biological parts that constitute the voice organ. Each of these parts plays a role in phonatory dynamics. An important goal of voice research is to understand how these parts and their interrelations

affect the dynamics of phonation. Without direct access to the control variables of the phonatory system and the equations that govern their evolution, this task becomes more demanding. This problem is partially addressed by using excised larynges, or in-vivo experiments with animals. In that case, the parameters can be experimentally controlled and thus the resulting dynamic behavior can be related to the control parameters. However, with excised larynges, the material properties are not exactly the same as in the living tissue, and much of the fine control of vocal fold posturing is lost.

1.5.3 Phase portrait reconstruction and the phasegram

Recently, a tool for visualizing phonatory dynamics from non-invasive measurements was developed by Herbst et al. (2013). This representation is named *phasegram* and it features the use of the electroglottogram for constructing a time history of phonatory states of a given voice production. Herbst et al. (2013) use the phasegram to visualize data that originates from models of the vocal folds, excised animal larynges and human phonation. The method utilizes a technique from nonlinear dynamical system analysis called phase portrait reconstruction (Crutchfield et al., 1980).

The phase portrait of a one-dimensional oscillator is a 2D graph that represents its motion, with one axis depicting its position and the other its velocity. The possible periodic oscillations are specific closed trajectories in this 2D space. To be able to plot such a phase portrait, the equations of motion of the system must be known. If the equations are not known, a phase portrait can be reconstructed based on a time series that originates from the system and is obtained from experimental measurements. For the phasegram, a phase portrait is reconstructed by having on the x-axis the EGG signal and on the y-axis either a delayed version of the EGG signal or its Hilbert transform. In case the Hilbert transform is used, the complex phase portrait y is constructed as the analytic signal of the EGG timeseries, i.e., the sum of the signal $\mathrm{EGG}[n]$ and its Hilbert transform $\mathcal{H}\{\mathrm{EGG}[n]\}$ multiplied with the imaginary unit i:

$$y = \mathrm{EGG}[n] + i\mathcal{H}\{\mathrm{EGG}[n]\}$$

An example of a phasegram alongside with a spectrogram and phase portrait displays can be seen in Fig. 1.1. The phase portraits are sampled at the intersection points with a straight line. These intersection points are called Poincaré sections and constitute the data that is used for constructing the phasegram.

1.5.4 Interpretation of the phasegram

The phasegram is very effective in visualizing phonatory dynamics and clearly demonstrates the transition between different vibratory regimes. However, as in the case of a spectrogram, the interpretation of a phasegram plot relies on the experimenter. Inspired by the phasegram approach, this work attempts to provide

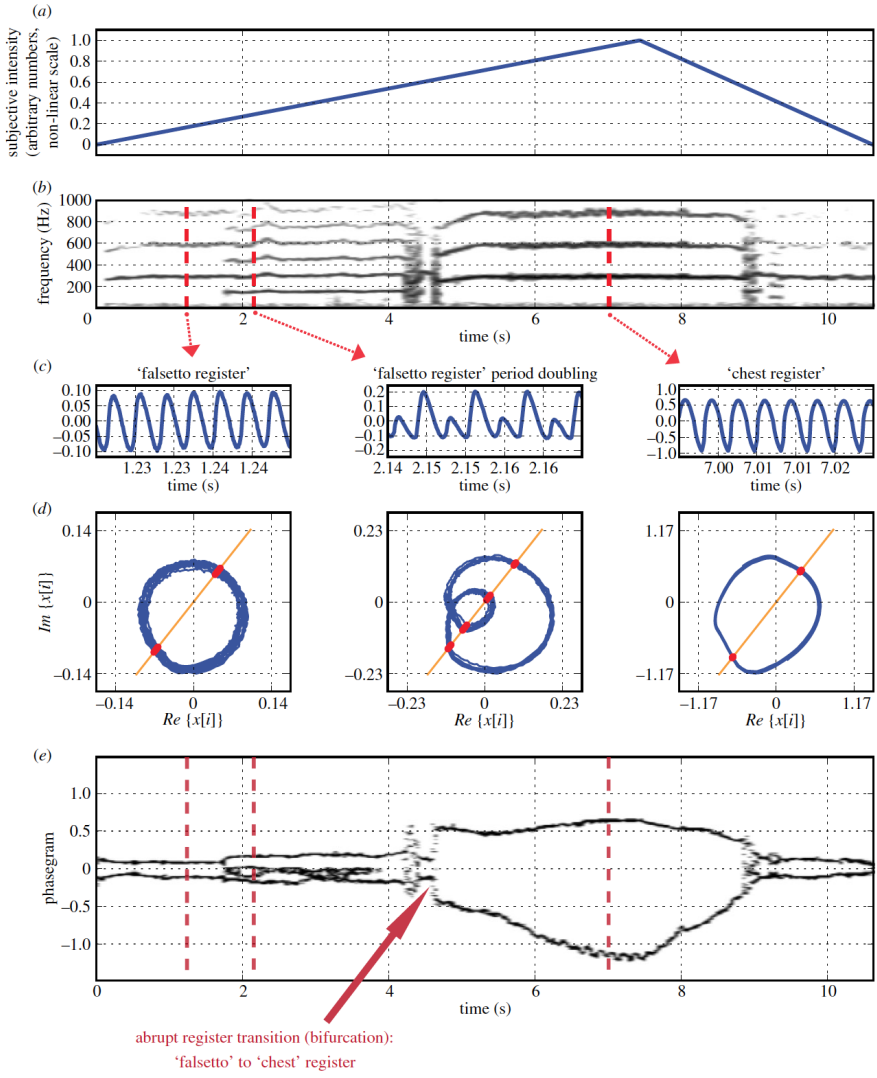

Figure 1.1: (From (Herbst et al., 2013), reprinted with permission of the author.) Electroglottographic recording of a 52-year-old singer producing a sustained note (vowel /a/) with [vocal] intensity variation (soft–loud–soft).(a) Singer's attempted intensity of voice production in arbitrary numbers on a nonlinear scale (0, lowest intensity; 1, highest intensity). (b) Narrow-band spectrogram of the electrographic (EGG) signal. (c) EGG signal, extracted at t=1.22, t=2.14 and t=7 sec. (d) Phase portraits from the above signals, created by plotting the real portion of the Hilbert-transformed EGG signal against its imaginary counterpart. A Poincaré section was created at an angle of 0.3π radians, yielding intersection points with the trajectory (red dots). (e) Phasegram: the vertical markers at t=1.22, t=2.14 and t=7 sec indicate the time instants at which the signals shown in panel (c) were extracted.

an extension towards the automated analysis of phonatory dynamics and states. It can be seen in Fig. 1.1d that different vibratory regimes are represented by characteristic phase portraits. Since the phasegrams are constructed based only on the Poincaré sections (red dots) of these phase portraits, some of the information on the vibratory state is lost. For example, the falsetto (leftmost in Fig. 1.1d) and chest (or modal) registers (rightmost in Fig. 1.1d) despite being different regimes of vibratory motion, are manifest in the phasegram (Fig. 1.1e from 0–2 sec and 4.5–9 sec) mostly as a change in the EGG amplitude. Typically in falsetto register, the contact area of the vocal folds is smaller than in modal register and as a result the EGG amplitude is reduced. Still, apart from the change in magnitude, the details of the phase portrait that characterize and differentiate the contact area patterns of falsetto and modal register are of interest, and not represented by the phasegram.

1.5.5 Short term analysis of vibratory states

A strategy for tackling the problem of detecting vibratory state changes and characterizing different vibratory regimes is the short-term analysis of the phase portrait. For detecting state changes, the phase-portrait can be analyzed cycle-by-cycle, using a stationarity measure that would exhibit a low value when the phase portrait is stable, then increase when a sudden change in the phase portrait occurs, and finally settle back to a low value as the phase portrait stabilizes again. Such a stationarity measure is discussed in Sec. 2.2. The cycle-by-cycle analysis has the advantage that it is adapted to the time window of interest; a disadvantage though is that it is not applicable in the most general case when the voice lacks periodicity, e.g., in cases of pathologies. Another interesting aspect of phonatory dynamics and states are smooth transitions, that is, transitions between different regimes that are not readily perceivable, or do not show as an abrupt change in the phase portrait. This type of transition is harder to detect, though it is expected that different regimes will show distinct characteristics that contribute to their identification.

The instability of the phase portrait may or may not be a result of a change in vibratory state. To assess whether an instability corresponds to a change in vibratory state, in addition to the stationarity measure, the phase portraits can be analyzed using some characteristic feature that describes their shape. A possible measure that can be used to characterize 2D closed curves is the Fourier descriptors. The Fourier descriptors are defined as the Discrete Fourier Transform (DFT) of the complex phase portrait. For the positive DFT bins, it can be shown that the DFT of any analytic signal, is equivalent to the DFT of its real part, in that case, the EGG signal. In effect, analysis of the phase portrait shape based on the DFT is the same procedure as pitch-synchronous spectral analysis of the one-dimensional EGG signal.

1.5.6 Vibratory states on a VRP

If an EGG based state differentiation scheme is coupled to a VRP measurement, the result is an enriched VRP that has the potential of revealing patterns on how

vibratory states dynamically interchange in the (F0, SPL) plane. The analysis could be applied to analyze how different paths in the VRP may affect the transition from one vibratory state to another.

1.6 Problem statement

The vocal folds, being a complex nonlinear oscillating system, are able to vibrate in a number of different ways, which are relevant both to professional voice users and clinicians. In the absence of data from invasive techniques, detecting the different vibratory states of the vocal folds is a demanding task. The purpose of this work has been to establish an analysis framework based on non-invasive techniques, to detect and discriminate between different vibratory states of the vocal folds, without prior knowledge on how these states manifest themselves in the measured signals.

2 Methods

2.1 Analysis of the electroglottogram

2.1.1 EGG analysis in the time-domain

Most commonly, EGG signals are analyzed in the time domain. A typical period of an EGG signal and its time derivative (dEGG) is shown in Fig. 2.1. Parameters that are used in EGG analysis are the EGG amplitude, the amplitude of the EGG derivative, the contact quotient and the rise time. The contact quotient (C_q), is defined as the fraction of the vibratory cycle that the vocal folds are in contact. The value of the C_q depends on the criterion used to define the contacting and decontacting events (Herbst and Ternström, 2006). Some researchers use threshold based values (e.g., define the contact and de-contact points of the cycle at the time instances when the amplitude of the EGG signal reaches some percentage of its maximum value) while others use the derivative of the EGG (dEGG) and its peaks. As seen in Fig. 2.1 the strong positive peak of the dEGG can be used to mark the contact instant and the –in this case weak– negative peak the decontacting instant. The rise time (RT) is defined as the fraction of the cycle during which vocal fold contact is increasing, i.e., when the dEGG is positive and larger than a given threshold.

2.1.2 EGG analysis in the frequency domain

A parameter that can be used to describe the EGG spectrum is its average spectrum slope (Libeaux, 2010). The power spectrum of several cycles of an EGG signal is shown in Fig. 2.2. It can be seen that the spectrum is harmonically structured and that there is a negative spectrum slope as the amplitude of the harmonics decreases monotonically with increasing frequency. A disadvantage of the spectrum slope is that it is a scalar measure which may poorly represent the details of the EGG spectrum.

This work utilizes short-term spectral analysis of the EGG signal. Briefly, the EGG signal is first separated into cycles, using an algorithm based on the analytic signal, and for each cycle the Discrete Fourier Transform (DFT) is calculated. As was mentioned in Sec. 1.5.3, this is equivalent with analyzing the DFT of the phase

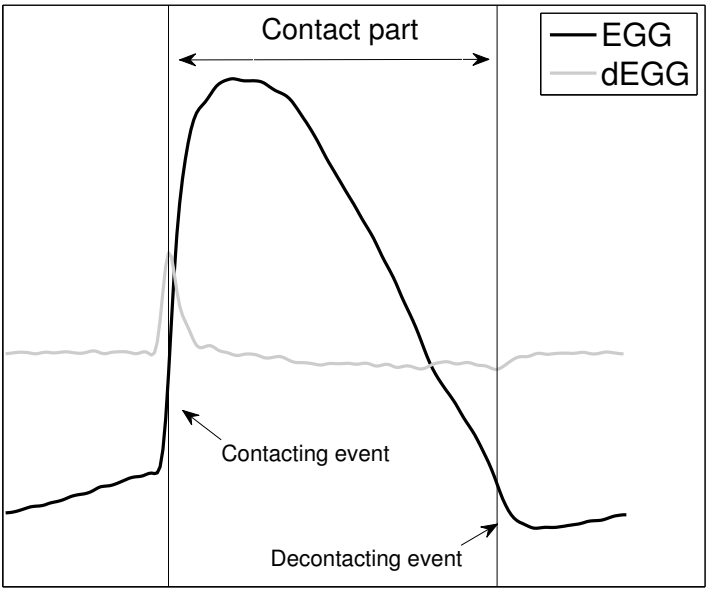

Figure 2.1: A period of an EGG signal. The black line corresponds to the EGG signal and the gray line to its time derivative.

portrait per cycle. This type of analysis is called pitch-synchronous and one cycle is regarded as the fundamental unit of vibration. With this approach, each bin of the DFT will correspond to a harmonic of the signal; thus it is straightforward to track the amplitudes and phases of individual harmonics. The pitch synchronous DFT components of the EGG signal are henceforth referred to as Fourier Descriptors (FD). Being interested mostly in the shape of the EGG curve, it is useful to normalize the DFT components relative to the fundamental. The relative levels ($\Delta L_n = L_n - L_1$) and phases $\Delta\phi_n = \phi_n - \phi_1$ of the harmonics will be referred to as 'Relative Fourier Descriptors' (RFD). Compared to the spectral tilt, this gives a finer description of the spectral distribution of the EGG cycle. The spectral tilt would coincide with the relative levels only in the case when the decay of the harmonics is uniform.

If the vibratory state of the vocal folds changes during a vocalization task, we expect that the EGG waveform will also vary in a way that reflects this change. Instead of analyzing the EGG waveform based on a few threshold points, the EGG spectral analysis can characterize the waveform of the entire cycle, still with a small number of data values. Another advantage of looking at the EGG signal in

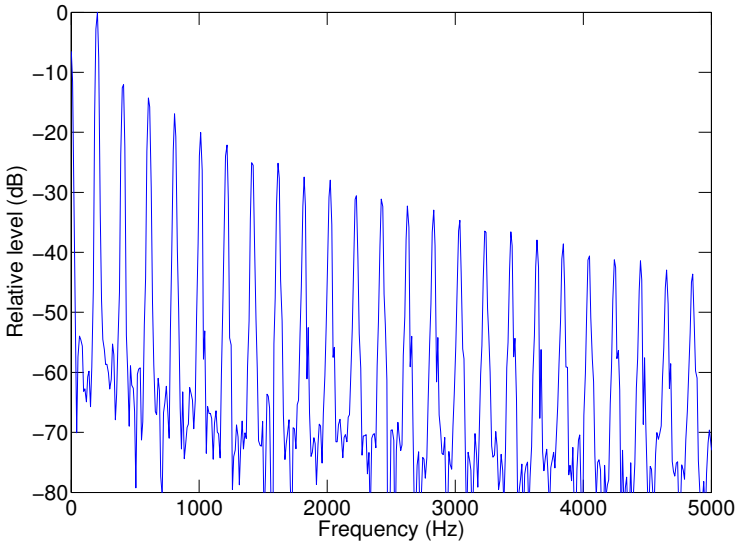

Figure 2.2: The power spectrum of several cycles of the EGG signal of a male subject phonating the vowel /a/ in modal phonation, at an F0 of 200 Hz.

the frequency domain is that one avoids having to define specific thresholds for contact. A typical problematic case may be such that there exist multiple peaks in the dEGG, when the dEGG based C_q becomes ill-defined. A disadvantage of the spectral approach though is that the Fourier components are not straightforward to interpret physically, in the absence of a model that would link different types of EGG waveforms to different biomechanical/vibratory conditions at the level of the vocal folds.

As an example of how the first Relative Fourier Descriptor relates to the contact quotient C_q a calculation with a simulated waveform was carried out. Figure 2.3 shows the results for a series of idealized half-rectified waveforms with different duty cycles (contact quotients) and the resulting level difference of the first two DFT harmonics. It can be seen that as the C_q increases, the first RFD ($\Delta L_2 = L_2 - L_1$) is decreasing. It should be mentioned however that the rise time of the EGG waveform also affects the spectrum slope and consequently the first RFD.

2.1.3 Limitations of EGG analysis

Though electroglottography as a method has many attractive features, there are some limitations worth mentioning. For example, the portion of the signal that corresponds to the open phase is not informative in any way about the state of the vocal folds. Additionally, in soft phonation when the vocal fold vibration amplitude is low, there might be little or no contact between the two vocal folds. In that case the analysis has to be complemented by examining other signals, e.g. the airborne

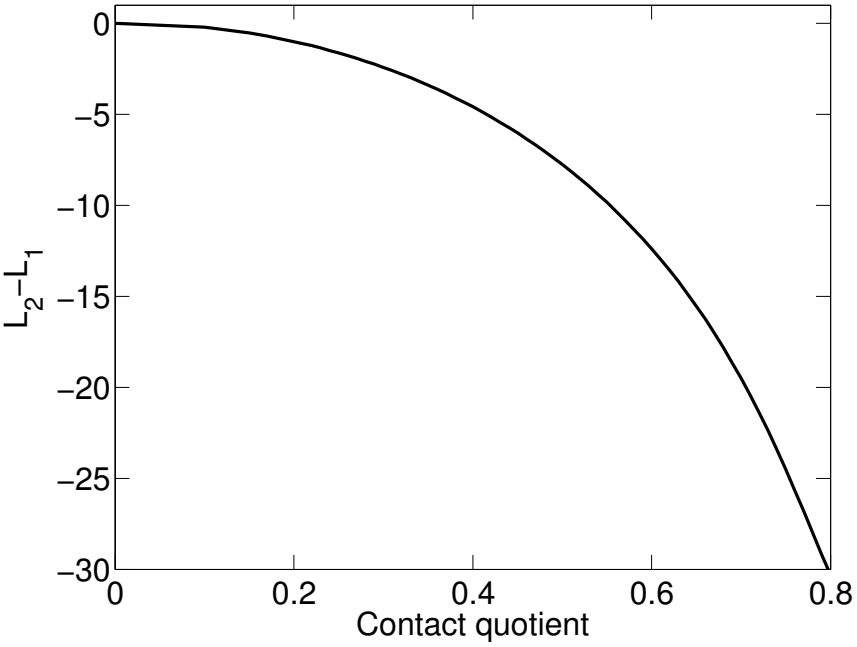

Figure 2.3: The contact quotient for synthesized half-rectified waveforms with different duty cycles. The level difference of the first two harmonics is shown as a function of the duty cycle.

signal or an accelerometer signal. Finally, the mapping between the EGG signal and specific mechanical aspects of vocal fold vibration is a difficult inverse problem that can be approximated only with the help of theoretical models for vocal fold motion.

2.2 Entropy measures in voice analysis

The need for methods to assess the complexity of biophysical time-series such as electrocardiograms (ECG) has led to the formulation of entropy measures like Approximate Entropy (ApEn) (Pincus, 1991) or Sample entropy (SampEn) (Richman and Moorman, 2000). These entropy measures estimate regularity by calculating the degree of self-similarity within a time-series. The great advantage of these tools is that they are robust against noise, and thus suitable for examining experimental data.

ApEn and more recently SampEn have been applied to EGG signals in clinical studies (Douglas et al., 2010; Fabris et al., 2013; Manickam et al., 2005), to quantify pathological phonation. These studies show how these measures are effective in

distinguishing between healthy and pathological groups of subjects and in general offer a more consistent interpretation of pathological signals compared to standard regularity measures like jitter (Fabris et al., 2013).

In this work SampEn was calculated for the pitch-synchronous time-series of the Fourier Descriptors, to detect the moments when the vocal folds changed their vibratory state. It was found that in an experimental task that contains vibratory state transitions, the SampEn exhibits a marked increase at the point/cycle of the transition. For details of how the SampEn was exploited in detecting state transitions please refer to Sec. III of the included paper.

2.3 Clustering using the Relative Fourier Descriptors

Clustering is a method to separate data into groups based on similarity. Clustering methods are widely applied in different disciplines that involve pattern recognition tasks. For an overview in methods used in pattern recognition and machine learning, see, e.g., (Bishop et al., 2006). A popular method used in clustering is the Gaussian Mixture Model (GMM). A GMM models the distribution of an input dataset along its dimensions as a sum of multidimensional Gaussian distributions. The parameters of the Gaussians (mean vectors and covariance matrices) are calculated in an iterative fashion using the EM (expectation maximization algorithm). The EM algorithm finds the set of parameters that maximize the posterior probability to observe t §he input dataset, given these parameters of the Gaussians.

Different types of EGG cycles can be grouped with regard to the corresponding vibratory state that generated them, by using the Relative Fourier Descriptor data as input to a GMM. As described in (Selamtzis and Ternström, submitted), a dataset containing two distinct vibratory states, namely modal and falsetto voice, was successfully clustered using a GMM based on the first four RFDs. A desirable property of this type of method is that it requires no prior knowledge about the different states that may exist in the dataset.

2.4 Outline of the proposed method

An overview of the proposed method is shown in the flowchart of Fig. 2.5. The experimental setup is shown in Fig. 2.4. The EGG signal and the laryngeal tracking output of the EGG device (Glottal Enterprises, EG-2, Syracuse, NY) are acquired synchronously with the airborne signal for a vocal production with state transitions at a sampling rate of 44.1 kHz. The A/D conversion and data acquisition is done by an acquisition board (model DT9834, Data translations Inc, Marlboro, MA) which connected via USB to a computer.

The EGG signal is high-pass filtered and separated into cycles based on the analytic signal, and for each cycle the DFT is calculated. The relative levels and phases of the DFT components per harmonic comprise the Relative Fourier Descriptors, i.e., the features that describe each EGG cycle. These features are used to group

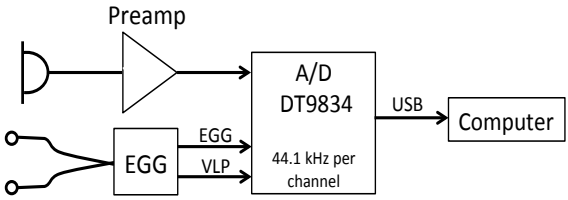

Figure 2.4: The schematic of the experimental setup.

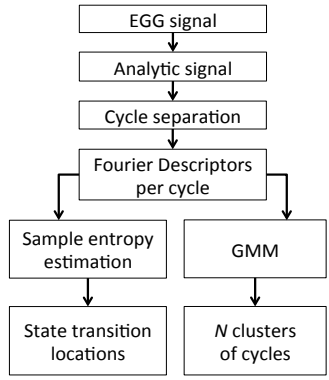

Figure 2.5: The processing routine for state transition detection and clustering.

the cycles that correspond to different vibratory states in clusters using a GMM and to locate state transitions by calculating SampEn and finding its maxima.

Figure 2.6 depicts the spectrogram, Fourier descriptors and combined Sample entropy for an upward pitch glide with a register break. It can be seen that at the moment of transition the SampEn presents a positive peak. The Fourier descriptors are also seen to abruptly change. The change of the Fourier descriptors is twofold; their amplitude uniformly diminishes by some factor, reflecting an abrupt decrease of the amplitude of the EGG signal, and at the same time there a different relationship between their levels and phases is established.

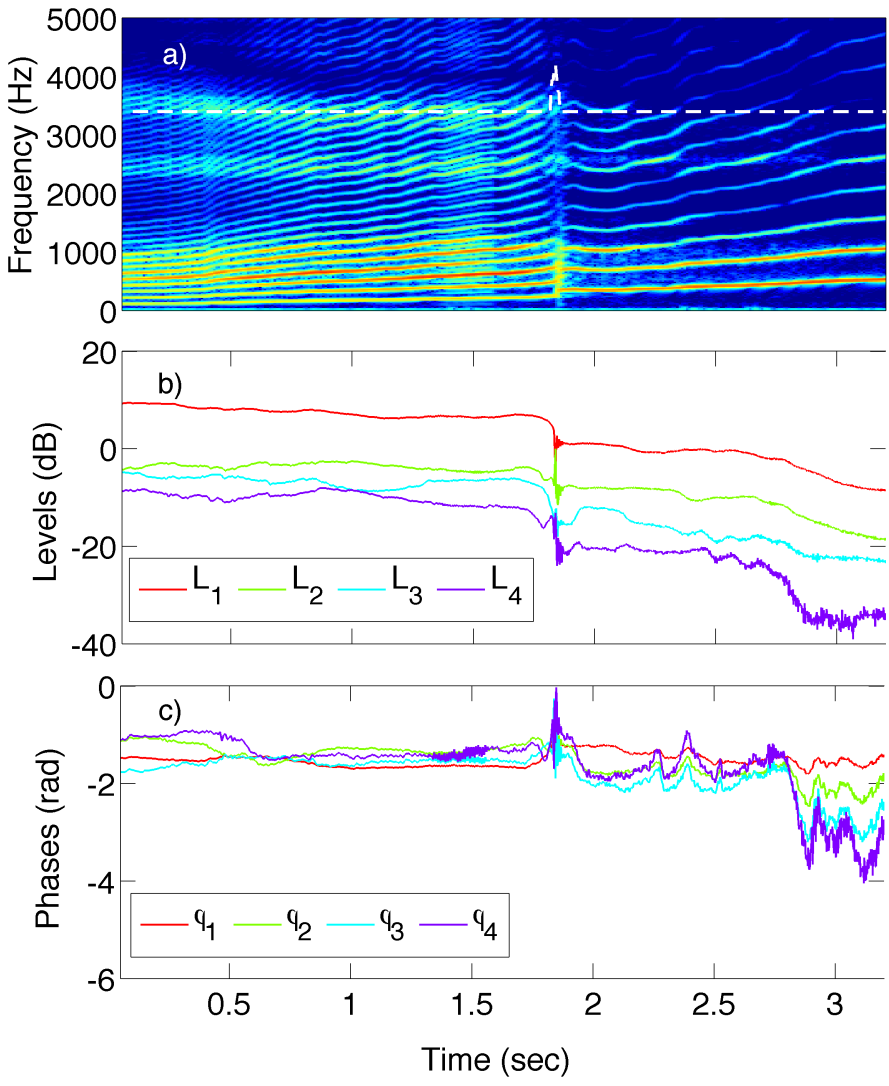

Figure 2.6: (a) The spectrogram for a pitch glide with one register transition. SampEn is the thick white dashed line in the spectrogram (arbitrary scale), (b) the levels of the Fourier descriptors (arbitrary reference), (c) The phases of the Fourier descriptors.

3 Results

3.1 Summary of results of the included paper

For the purpose testing the detection vibratory state changes, the SampEn was compared with manual annotations on spectrograms, for 50 EGG recordings containing register breaks. The 96% of the manually annotated transitions corresponded to peaks in SampEn, which shows the utility of it as a state detection indicator.

For categorizing different vibratory states based on the Relative Fourier Descriptors, amateur singer subjects phonated upward and downward pitch glides. The first three Relative Fourier Descriptors were extracted and given as an input to a GMM. The GMM separated the dataset with regard to the vibratory state into minimally overlapping groups (clusters), showing that the spectral features of the EGG suffice for identifying two distinct ways of vibration. The corresponding regions were mapped onto the VRP plot using the synchronous (F0, SPL) data that was extracted from the acoustic signal. The EGG-based clusters of data points were found to appear in connected regions of the (F0, SPL) plane. By further increasing the number of desired data clusters of the GMM the trend it was observed that the boundary between clusters that divided the modal region into sub-regions tends to be an almost horizontal line on the (F0, SPL) plane, while in the falsetto region the separation was along a diagonal line.

3.2 Sources of variation in the EGG

3.2.1 Background

By 'vocal fold posturing' we mean the positioning and pretensioning of the vocal folds, which is crucial for the resulting type of phonatory vibration. However, it is mostly executed by the small intrinsic laryngeal muscles, and therefore requires invasive methods for observation, such as electromyography. By 'laryngeal posturing' we mean the positioning of the laryngeal cartilages relative to each other, which sets up the constraints for vocal fold posturing; and also the vertical laryngeal position (VLP), i.e., the height in the neck of the entire larynx assembly, relative to a rest position. The VLP in particular is of great interest to practitioners. It is directly observable from the outside in most individuals, and with a little practice, it can

fairly easily be brought under conscious control. Because of the general mechanical interconnectedness of the components of the larynx, we may expect that variations in VLP will give rise to fluctuations also in the vibratory pattern of the vocal folds. In fact, the VLP may well have an influence on the tendency at any given moment of the vocal folds to change to another vibratory state.

It is beyond the scope of this study to investigate the influence of the VLP in detail, but we need some indication of whether or not the VLP could be a confounding factor for the automatic clustering of the vibratory states. It is therefore of interest to examine the nature of the covariation, if any, between the VLP and the FD components of the EGG signal. The dual-channel EGG device used for the present study has also an output whose voltage tracks the VLP (on an arbitrary scale), and this signal was acquired in parallel.

3.2.2 Tracheal pull and lung volume

The lung volume decreases monotonically during a vocal production. Past studies have shown that the lung volume can affect the glottal flow waveform (Iwarsson et al., 1998). The lung volume is linked to the tracheal pull, a downward force that may also affect the posturing of the vocal folds. The tracheal pull is considered to exert an abductory force on the vocal folds, bringing them slightly apart, when the lungs are full. The tracheal pull effect becomes more apparent in extreme cases, e.g., when the subject has inhaled maximally.

The relative lung volume can be measured by measuring the periphery of the ribcage. For the purposes of this study, no measure of the lung volume was acquired, but if other factors are constant we can see the effect of tracheal pull on the VLP. Figure 3.1 shows a sung production of a *messa di voce* exercise, i.e. a gradual increase and then decrease of vocal loudness, with the corresponding VLP trace in the lowest panel. The F0 is essentially constant and the singer stays in the modal register. The VLP is seen to increase throughout the production, especially near the beginning and near the end. This should be mostly an effect of the decreasing lung volume. It may be noted that the EGG FDs are relatively stable as the VLP changes.

3.2.3 Vertical laryngeal position

Figure 3.2 shows instead an upward glissando. Initially the VLP rises with F0, as has been observed by others for glissando tasks, while at the very top of the falsetto range, it reverses and drops a little. This behavior was very common. Note that while the EGG FDs are changing somewhat, these changes do not seem to be immediately linked to the VLP. Rather, the VLP changes gradually, and at certain points the EGG FD values that represent the vibratory pattern of the vocal folds change more abruptly.

Figure 3.3 shows an example of a subject switching between modal and falsetto, while trying to maintain the same F0. Here the alternation of FD patterns is quite

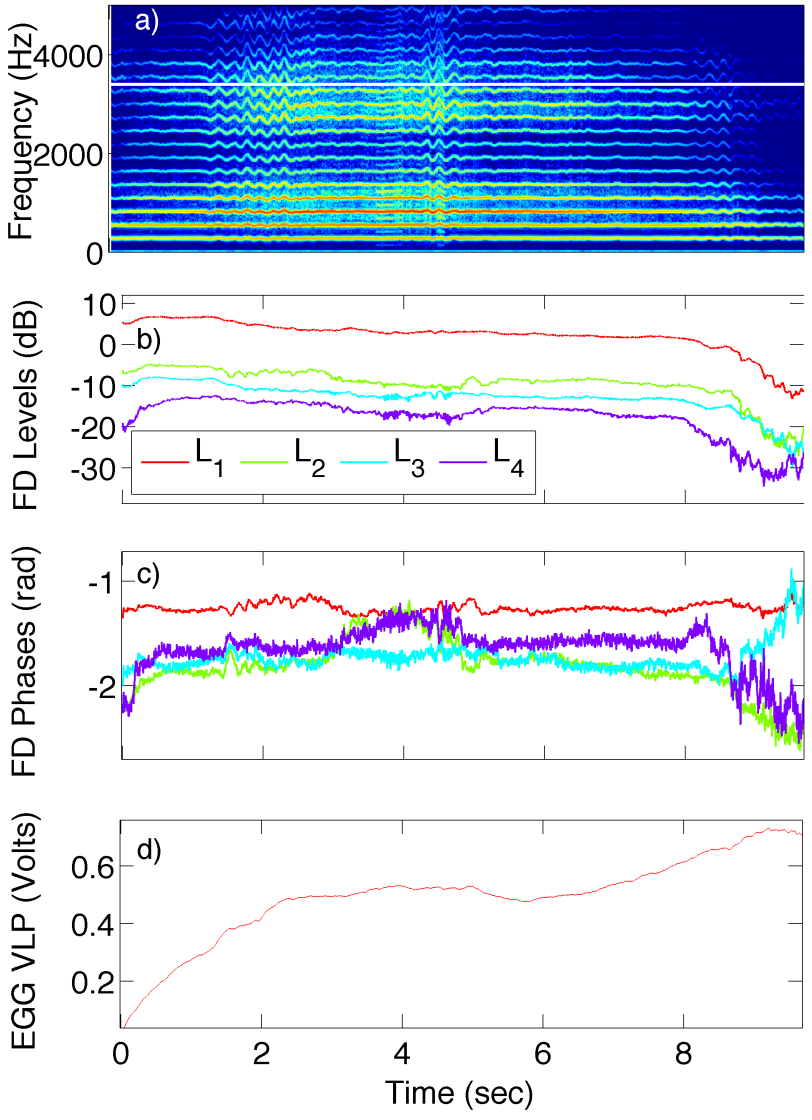

Figure 3.1: (a) The spectrogram of a *messa di voce* (*crescendo* and *decrescendo* on the same F0), (b) the corresponding FD levels, (c) the corresponding FD phases, (d) the indication of VLP taken by the "Laryngeal Tracking" output of the EGG device (arbitrary scale). The white line in (a) depicts the SampEn (arbitrary scale).

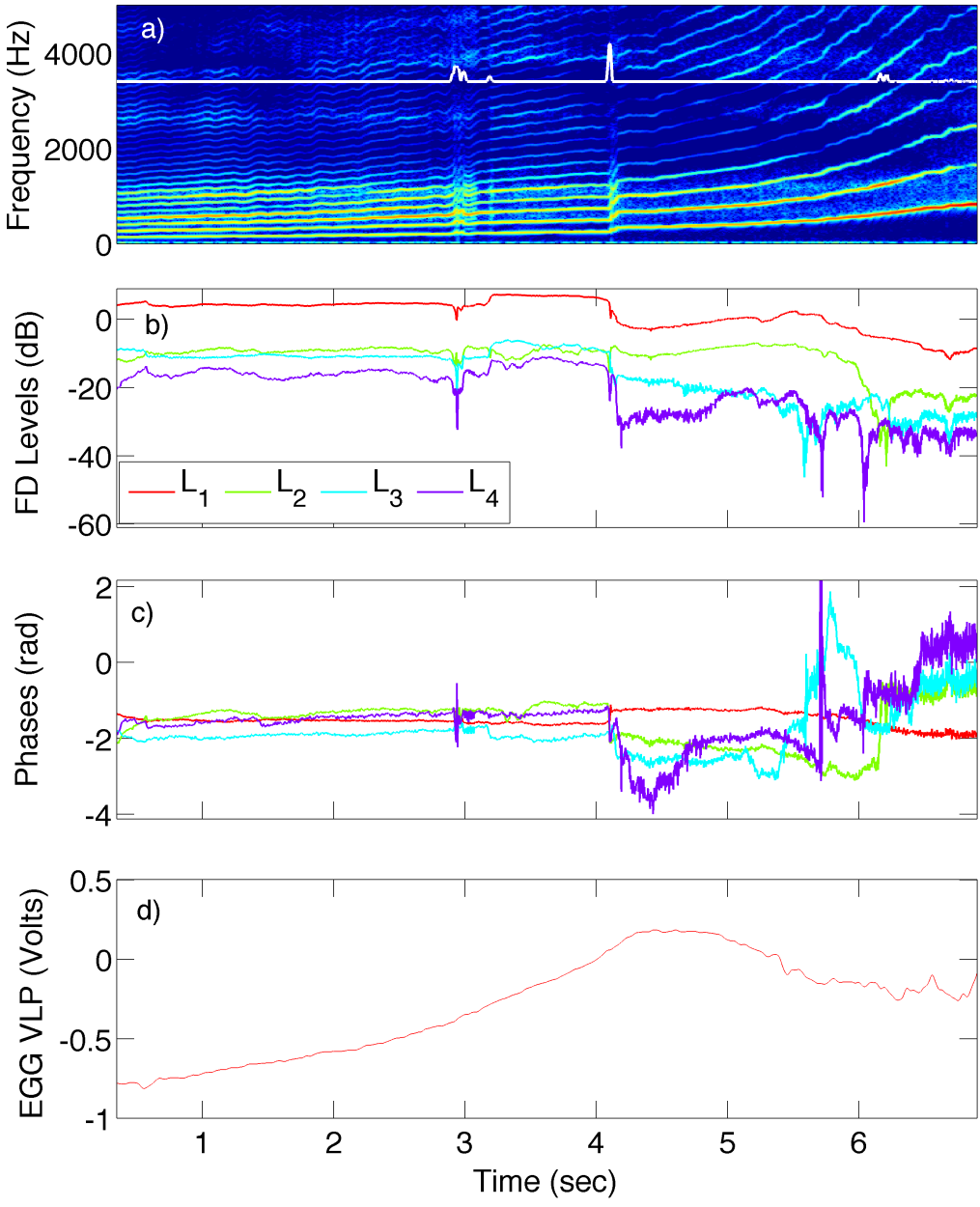

Figure 3.2: (a) The spectrogram of pitch glide an F0 jump and a register break, (b) the corresponding FD levels, (c) the corresponding FD phases, (d) the indication of VLP taken by the "Laryngeal Tracking" output of the EGG device (arbitrary scale). The white line in (a) depicts the SampEn (arbitrary scale).

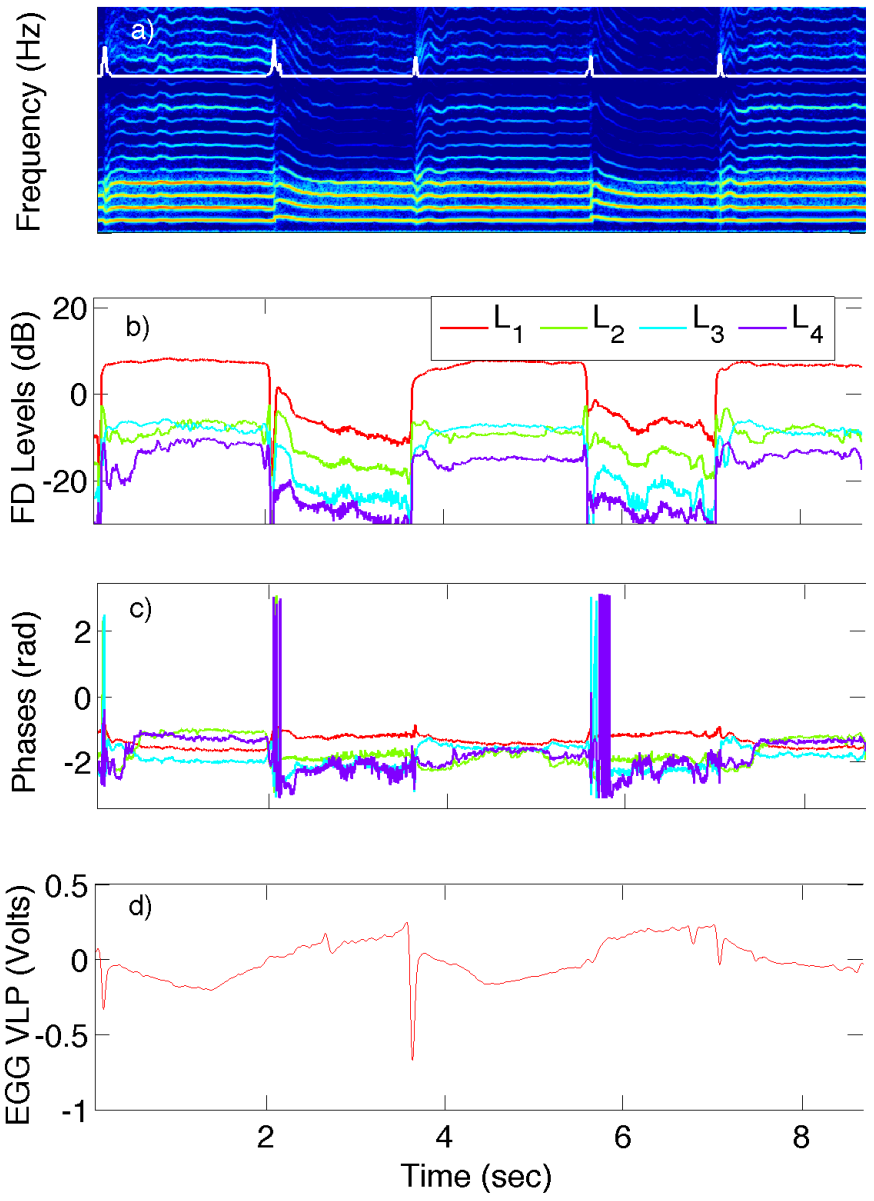

Figure 3.3: (a) The spectrogram of a number of modal/falsetto transitions while keeping F0 constant, (b) the corresponding FD levels, (c) the corresponding FD phases, (d) the indication of VLP taken by the "Laryngeal Tracking" output of the EGG device (arbitrary scale). The white line in (a) depicts the SampEn (arbitrary scale). The vertical transients in (c) are modulo-2π wrap-around artefacts which are not factored into the SampEn measure.

abrupt; but, while the VLP does covary with vibratory state, it clearly is influenced by other factors as well, such as changing tracheal pull. Some but not all of the state transitions are accompanied by a transient in the VLP.

3.2.4 Effect of vowel

In the experiments conducted for the purpose of the included paper, the vowel /a/ was used in all recordings, and it was kept constant during the whole production. However, some changes in the Fourier descriptors of the EGG can be seen as the vowel changes. Figure 3.4 depicts the spectrogram and the FDs levels and phases for a production with five vowels /a/, /e/, /i/, /o/, /u/. Despite the fact that the FDs do vary, the SampEn curve (white dashed line in the spectrogram) does not show any variation. It is instructive to compare Fig. 3.4 with Fig. 2.6. A visual comparison on the scale suggests that the variation of the FDs in the case of vowel transitions is much smaller compared to the variation they exhibit in cases of register transitions. That is also indicated by the behavior of the SampEn; in the case of Fig. 3.4, SampEn remains constant while in the case of the register transition it increases markedly at the instant of the transition. It should be noted that the behavior of the SampEn depends on its adjusted sensitivity, so this involves a somewhat arbitrary factor, however after experimentation the adjustment can be tuned to the scale of the desired phenomenon (e.g., a register transition or a vowel change).

When listening to the EGG signal, the differences between vowels were mostly inaudible. Still, even with these small variations in the FDs, the GMM clustering was able to separate the dataset into four clusters with */i/* and */u/* being assigned in the same cluster. This does not necessarily indicate that source-filter interaction is in play. Even changing the articulation from one vowel to another could slightly affect the laryngeal posturing.

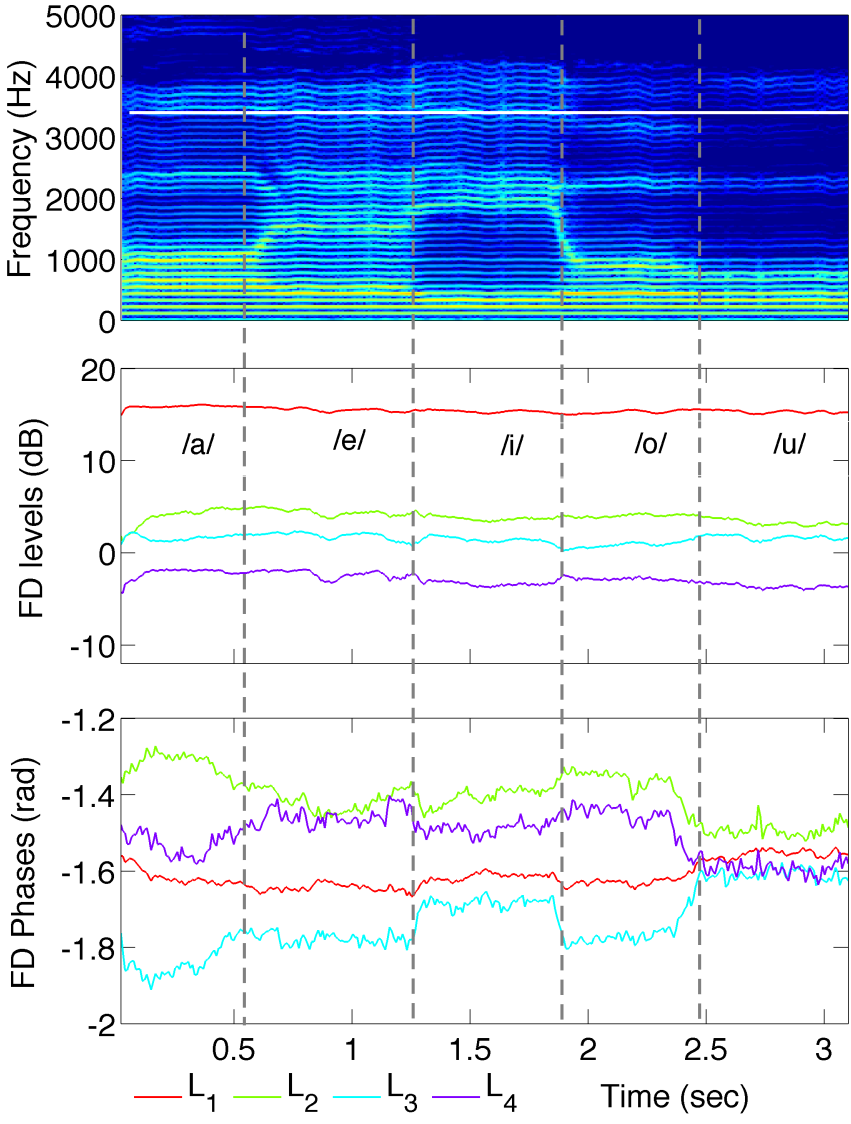

Figure 3.4: (a) The spectrogram for a sustained tone on a sequence of 5 different vowels. SampEn is the thick white dashed line in the spectrogram (arbitrary scale), (b) the levels of the FDs, (c) The phases of the FDs.

4 Discussion

The aim of this work has been twofold: First, to explore the utility of the short-term spectrum of the EGG signal as an indicator of the dynamics of vocal fold behavior, and secondly to provide a tool for automatically marking instability regions of the voice as well as segregating between different vibratory states. A representation based on the proposed method can enhance the VRP and increase its potential as a research tool for the dynamics of the voice. A pedagogical application of this could be the interactive exploration of the limits of operation for different vibratory states, or voice registers. By means of depicting high SampEn values on the VRP, regions of instability could be also interactively identified. For pre-contacting phonation, when the EGG is very weak, the EGG analysis could be complemented with accelerometer signals.

The approach using clustering algorithms need not be restricted to the Relative Fourier Descriptors of the EGG. This approach could work with any other waveform derived from a physiological measurement that relates to the vocal fold vibration, such as for example the glottal flow from inverse filtering, the glottal area, or photoglottographic waveforms. It can be expected that specific patterns will again show up that will segregate different vibratory states.

Further interpretation of the correspondence between the differences in the waveform and the corresponding physiological phenomena could be obtained by using models of the EGG waveform. Additionally, more invasive data could be used in conjunction with the proposed non-invasive methods to assess whether subtle differences that show up by the GMM algorithm as different subregions within vibratory states originate from different phenomena on the laryngeal vibratory level.

For a future clinical application, the method could be used to find the regions of F0 and SPL where the subject's voice is least unstable. However, pathological signals may be excessively irregular and that might be a serious limiting factor for the clustering algorithm.

The research effort described in this report establishes a framework for further investigation of phonatory dynamics. Aspects that will extend this work in the future include the extension of the method to include acoustic analysis, e.g., formant tracking to investigate possible source-filter interactions, inclusion of protocols that utilize phonation in different degrees of glottal adduction for assessing how adduc-

tion may affect the vibratory states and the dynamic transitions between them, and investigation for applications in the clinical domain.

Bibliography

Baken, R. J. (1992). Electroglottography. *Journal of Voice*, 6(2):98–110.

Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 1. Springer New York.

Crutchfield, J., Farmer, J., Packard, N., and Shaw, R. (1980). Geometry from a time series. *Phys. Rev. Lett*, 45(9):712–716.

Davis, P. J. and Fletcher, N. H. (1996). *Vocal fold physiology: Controlling complexity and chaos*. Singular Pub Group.

Douglas, C. M., Moore, C., Manickam, K., Lee, L., Sykes, a., Carr, a., Jones, S., Jones, J., Swindell, R., Homer, J. J., and Slevin, N. (2010). Electroglottogram approximate entropy: a novel single parameter for objective voice assessment. *The Journal of laryngology and otology*, 124(5):520–8.

Fabris, C., De Colle, W., and Sparacino, G. (2013). Voice disorders assessed by (cross-) Sample Entropy of electroglottogram and microphone signals. *Biomedical Signal Processing and Control*, 8(6):920–926.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton, The Netherlands.

Henrich, N., D'Alessandro, C., Doval, B., and Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America*, 115(3):1321.

Herbst, C. and Ternström, S. (2006). A comparison of different methods to measure the EGG contact quotient. *Logopedics, phoniatrics, vocology*, 31(3):126–38.

Herbst, C. T., Fitch, W. T. S., and Švec, J. G. (2010). Electroglottographic wavegrams: a technique for visualizing vocal fold dynamics noninvasively. *The Journal of the Acoustical Society of America*, 128(5):3070–8.

Herbst, C. T., Herzel, H., Švec, J. G., Wyman, M. T., and Fitch, W. T. (2013). Visualization of system dynamics using phasegrams. *Journal of The Royal Society Interface*, 10(85):20130288.

Herzel, H., Berry, D., Titze, I., and Steinecke, I. (1995). Nonlinear dynamics of the voice: signal analysis and biomechanical modeling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(May 2011):30.

Iwarsson, J., Thomasson, M., and Sundberg, J. (1998). Effects of lung volume on the glottal voice source. *Journal of Voice*, 12(4):424–433.

Jiang, J. J., Zhang, Y., and McGilligan, C. (2006). Chaos in voice, from modeling to measurement. *Journal of voice : official journal of the Voice Foundation*, 20(1):2–17.

Lamarche, A. (2009). *Putting the Singing Voice on the Map: Towards Improving the Quantitative Evaluation of Voice Status in Professional Female Singers*. PhD thesis, KTH.

Libeaux, A. (2010). The egg spectrum slope in speakers and singers: Variations related to voice sound pressure level, vowel and fundamental frequency. Master's thesis, KTH.

Lucero, J. C. (1996). Chest-and falsetto-like oscillations in a two-mass model of the vocal folds. *The Journal of the Acoustical Society of America*, 100(5):3355–3359.

Lucero, J. C. (2005). Oscillation hysteresis in a two-mass model of the vocal folds. *Journal of Sound and Vibration*, 282(3-5):1247–1254.

Manickam, K., Moore, C., Willard, T., and Slevin, N. (2005). Quantifying aberrant phonation using approximate entropy in electrolaryngography. *Speech Communication*, 47(3):312–321.

Mehta, D. D., Zanartu, M., Feng, S. W., Cheyne, H. A., and Hillman, R. E. (2012). Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *Biomedical Engineering, IEEE Transactions on*, 59(11):3090–3096.

Neubauer, J., Edgerton, M., and Herzel, H. (2004). Nonlinear phenomena in contemporary vocal music. *Journal of voice : official journal of the Voice Foundation*, 18(1):1–12.

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301.

Plant, R. L., Freed, G. L., and Plant, R. E. (2004). Direct measurement of onset and offset phonation threshold pressure in normal subjects. *The Journal of the Acoustical Society of America*, 116(6):3640.

Ptok, M., Sesterhenn, G., and Arold, R. (1992). [evaluation of laryngeal sound generation with fft analysis of glottic impedance in patients with recurrent nerve paralysis]. *Folia phoniatrica*, 45(4):182–197.

Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology. Heart and circulatory physiology*, 278(6):H2039–49.

Selamtzis, A. and Ternström, S. (submitted). Analysis of vibratory states in phonation using spectral features of the electroglottographic signal. *The Journal of the Acoustical Society of America*.

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A. M., Tyack, P., Samarra, F., Miller, P., and Wallin, J. (2014). Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America*, 135(2):953–962.

Strogatz, S. H. (2006). *Nonlinear dynamics and chaos (with applications to physics, biology, chemistry and engineering*. Perseus Publishing.

Švec, J. G., Schutte, H. K., and Miller, D. G. (1999). On pitch jumps between chest and falsetto registers in voice: data from living and excised human larynges. *The Journal of the Acoustical Society of America*, 106(3 Pt 1):1523–31.