Electroglottography in Real-Time Feedback for Healthy Singing

ISAK NILSSON
Master Thesis Report

of Isak Nilsson (isakni@kth.se)

Royal Institute of Technology, Sweden

Department of Speech, Music and Hearing

Supervised by Prof. Sten Ternström

Examiner: Prof. Olov Engwall

October 28, 2016
Electroglottography in Real-Time Feedback for Healthy Singing

Abstract

This master thesis describes early attempts at using electroglottography (EGG) to capture such changes in vocal fold vibration patterns that could be of interest to teachers of contemporary commercial music. After initial explorations, focus is placed on detecting potentially detrimental phonation; more specifically on the pressed quality often associated with loud singing in high register (belting). FONADYN, a program written in the SuperCollider language, is used to detect pressedness using an algorithm based on K-means clustering of Fourier components of EGG cycles. Results indicate that pressedness affects phonation in ways detectable using EGG. Changes caused by pressedness seem to vary between registers and this variation is similar between subjects. Detection of pressedness in a subject is quite successful when training the algorithm on the same subject, but not always across subjects.

Elektroglottografi i realtidsfeedback för hållbar sångteknik

Sammanfattning

Denna masteruppsats beskriver inledande försök att använda elektroglottografi (EGG) för att avläsa sådana förändringar i stämbandens vibrationsmönster som skulle kunna vara av intresse för sånglärare inom icke-klassisk stil. Tidiga undersökningar leder till att fortsatt fokus läggs på att detektera fonationstyper som kan orsaka röstskador; mer specifikt den typ av pressad röstkvalitet som ofta förknippas med stark sång i högt register (s.k. belting). FONADYN, ett datorprogram skrivet i språket SUPERCOLLIDER, används för att detektera pressad fonation med hjälp av K-means-klustring av EGG-cykler baserat på deras Fourierkomponenter. Resultaten indikerar att pressad fonation går att urskilja med hjälp av EGG. Kännetecknen för pressad fonation tycks skilja sig mellan röstregister och denna skillnad är snarlik hos olika försökspersoner. Programmet klarar av att känna igen pressad fonation hos samma person som algoritmen tränats på men inte alltid om algoritmen tränats på en annan sångare.
Contents

1 Introduction 1

2 Background 1
   2.1 Contemporary Commercial Music 1
   2.2 Real-Time Singing Feedback 3
   2.3 The Voice Range Profile 4
   2.4 Electroglottography 6
   2.5 The FONAdyn Software 7

3 Method 10
   3.1 Identifying Relevant Features 10
      3.1.1 Feature Candidates 10
      3.1.2 Selection Process 10
   3.2 Recording Setup and Protocol 11
      3.2.1 Hardware Setup and Calibration 11
      3.2.2 Software Settings 12
      3.2.3 Recording Protocol 13
   3.3 Data Processing 14
      3.3.1 Constructing Training and Evaluation Material 15
      3.3.2 From Training Material to Clusters 16
      3.3.3 From Clusters to Evaluated Cycles 18
      3.3.4 Analysis of Clusters as Vectors 19

4 Results 20
   4.1 Analysis of Clusters as Vectors 20
      4.1.1 Distances between Cluster Centroids 20
      4.1.2 Directions of Increasing Pressedness 21
      4.1.3 Comparison of Fourier Descriptors 23
   4.2 Accuracy of Cluster-based Classification 24
      4.2.1 High-pitched Recordings 26
      4.2.2 Low-pitched Recordings 27
      4.2.3 Long-term Consistency 28

5 Discussion 29
   5.1 Normal and Pressed Quality Cycle Shapes 29
   5.2 Cluster Properties and Evaluation Accuracies 31
   5.3 Discussion of Sustainability 33

6 Concluding Remarks 34

7 Acknowledgements 34

References 35
1 Introduction

Since its introduction in 1956, electroglottography (EGG) has been found to be a useful yet non-invasive way to monitor phonation. With increasing processor capacities, it has become possible to use EGG measurements for real-time feedback on almost any PC machine. This opens up new possibilities for the use of EGG measurements outside of research laboratories. For instance, by giving objective and immediate feedback, use of EGG could remedy the ambiguity and delay problems commonly observed in singing teaching situations - provided that the relevant features can be reliably captured by EGG.

This report describes early attempts to identify and capture phonation features of contemporary commercial music, using software recently developed for the purpose of real-time analysis and display of EGG data, as a first step towards the long-term goal of making EGG a useful and readily available tool in teaching singing. The project is an international collaboration between engineers and singing pedagogues, conducted at the Royal Institute of Technology in Sweden.

Below, the Background section introduces relevant concepts and the specialised FONADYN software. Next, the stages of identifying relevant features and collecting and treating EGG data are outlined. Finally, comparisons of data between multiple subjects and for a single subject over time are made, concluded by some remarks on the challenges ahead.

2 Background

2.1 Contemporary Commercial Music

Contemporary commercial music (CCM) is a recently introduced term used to refer to a wide range of non-classical music genres. It includes among other genres jazz, pop, rock, soul, blues and musical theatre (see Fig. 1), genres in which the singing ideals and techniques differ from those of classical music [1].

In CCM, the singing more closely resembles speech in the sense that vowels comprise a smaller percentage of phonation as compared to classical singing; the CCM singing style is less legato and more rhythm-driven. It has also been said that in classical singing, there exist well-defined singing ideals and techniques for which to strive in order to achieve expected pitch ranges, sound pressure levels and timbre.
Contemporary Commercial Music

Gospel  Country  Jazz
Pop  Music Theatre  Rock
Folk  Alternative

Figure 1: Contemporary commercial music (CCM) is a collective term including many non-classical genres.

One thing that sets CCM apart from classical singing is electric amplification. The use of microphones enables many new expressive voice qualities to be used, since the singer’s formant is no longer necessary to reach a large audience or fill a concert hall with song [1, 2]. CCM voices thus have lesser need for acoustical power, allowing for greater disparity in voice qualities between artists. Being unique and recognisable has proven useful in becoming famous and prosperous. Many successful CCM singers thus make an effort to create and maintain their very own unique sound, using alternative techniques and avoiding classical voice ideals. There are also artists who instead of producing original material make their living by impersonating others, striving to make live performances that come as close as possible to recordings of the original artist. Such cover artists often mix songs from different original artists, all the while adapting to mimic their respective voice characteristics [1]. Similarly, in musical theatre the singer must often alter their voice characteristics to portray a character [2]. It stands to reason that different techniques are necessary for classical and CCM singing.

The classical voice training does not explicitly include methods for controlling the highly specialised CCM techniques. Whether or not it indirectly gives the necessary preparation has been debated by students and teachers alike. Many singing teachers agree that the classical method is the only viable option, and that it sufficiently prepares the student for any type of singing. If the student is later harmed by CCM singing, this is not considered to be caused by inadequacy of the training but rather blamed on damage build-up inherent in CCM [1]. However, there exist CCM singers with very long and vocally healthy careers behind them, indicating that CCM and voice damage are not synonymous and perhaps not causally correlated at all [1, 2].
Possibly because CCM singing technique has been presumed to be indirectly covered by classical training, relatively few studies have been made which focus on the healthy CCM singing voice. There are voice pathology studies including CCM singers, but the subject necessarily means that damage has already been done. Some studies have indicated that young CCM singers may be at greater risk of developing voice issues, as are all female singers, but the subject remains relatively unexplored [2].

2.2 Real-Time Singing Feedback

Singing is a complex activity with many parameters to control; teaching singing techniques to others is correspondingly complex. To transfer knowledge, a combination of explanation and imitation is used. When explaining phonation techniques, possible issues include imprecise vocabulary (metaphors and psychological hooks), limited anatomical knowledge (in both student and teacher) and miscommunication and misunderstanding in general. The main issues with imitation include the possibility that very similar sounds could be produced employing different mechanisms or technique (healthy or damaging), and the assumption that the student already has sufficient aural skills to tell whether the attempt sounds right or wrong. Common to both alternatives is the fact that student attempts and teacher feedback are typically separate in time and thus introduce a delay in the attempt-feedback-correction cycle (see Fig. 2) [3, 4].

![Diagram](Fig. 2: a) Usually, a student attempting to perform a task will receive feedback after the attempt is done, introducing a delay between attempt and feedback, and between consecutive attempts. b) With real-time feedback, attempts can be made continuously since feedback is given in parallel. (Freely after [3])

With the advent of increasingly fast computers, efforts have been made to create real-time feedback systems that significantly shorten this cycle [4, 5]. Preferably, the student should receive feedback fast and often enough to experience the response as instant and continuous (see Fig. 2). Several software applications have been developed over the years, exploring different options and utilising the growing memory and processor capacities of
home computers [3, 5, 6]. These attempts have focused on improving a student’s pitch control and/or timbre. Typically, the acoustical signal has been used and the goal has been to produce a pleasant sounding voice; the vocal health of the student has not been in focus. Among the metrics that have been presented in different feedback schemes are pitch, vocal tract area, voice spectrum, formant resonances, vocal identity, and closed quotient [5]. The kind of feedback has also varied but is typically visual, either giving the full details and history of a metric or a binary indication of whether or not the metric currently falls within the target zone [6]. The reception has generally been positive, though teachers have had different opinions about how such tools are best used and tailored to students with different levels of training. It is generally agreed upon that teacher supervision is necessary when using the tool, and that the teacher’s role is not diminished by the use of such feedback systems [3–6].

2.3 The Voice Range Profile

The voice range profile (VRP) is a scheme for visualising the vocal range of a singer in a two-dimensional plot [7, 8]. The horizontal axis usually shows tempered semitones or MIDI numbers and the vertical axis shows sound pressure level (SPL) in decibels (dB). This results in a grid of cells, each corresponding to a certain pitch and volume (see Fig. 3). The x axis thus goes from low pitch to high pitch whereas the y axis goes from low sound level to high sound level (from quiet to loud). A singer will be able to reach some of these cells, but not all. Some pitches are too low or too high, and these limits vary with dynamic level; the highest pitches cannot be sung softly, and the very lowest notes cannot be sung loudly. In the VRP, the cells a singer has reached are filled in, while the yet unreached (and the altogether unreachable) remain blank. The result is usually an oblong shape centred at the comfortable singing range of the subject being recorded, tilted along the diagonal so that higher pitches are seen for louder sound levels and vice versa. Instead of simply marking the reached cells, a grey scale or colour gradient can be used to plot some interesting metric, such as accumulated time spent in the cell, spectrum richness of the voice and many other options [8–10]. Typically, this metric has been derived from the acoustical signal, but any metric could be used as long as it gives meaningful information about the voice (see Fig. 3).

A VRP clearly shows the voice range of a singer and for healthy singers typically consists of a single connected oblong shape [8–10]. Most VRP plots share common features of the contour; a dip in the upper contour is often seen at the border between modal and falsetto phonation, since it is harder to sing loudly in that area; an upwards sloping section of the upper contour has been found to be very consistent between subjects (neither of these
Figure 3: Examples of a voice range profile (VRP) for a female singer (staying in a CCM style, avoiding classic style) for four different choices of the third metric. Top left: Clarity. Cycles are rejected (grey) or kept for analysis (green) based on some threshold value. Top right: Accumulated time. Bottom left: Crest factor. Bottom right: Cluster-based classification. Colours represent different cycle shapes.

is shown in Fig. 3) [8]. However, even VRP plots produced by the same equipment and subject can differ over time, for instance because of voice training [9].

Before measuring the VRP of a subject, several things must be taken into consideration. Microphone configurations, computational algorithms and possibly the third metric to be used must be chosen in appropriate ways [8]. For instance, the kind of microphone and its placement will affect what levels of sound pressure can be reliably recorded. Using a boom microphone mounted in a headset, close to the mouth, will probably allow for separating faint singing from background noise compared to when using a microphone mounted on a stand further from the singer. The extra sensitivity to faint sounds will extend the lower contour of the VRP (as in Fig. 3 where the VRP extends all the way down to the edge of each plot). Such differences must be taken into account when comparing different VRP recordings, using different equipment, arrangements or protocols [8, 9]. The algorithms used for processing the recorded data should also be chosen in a way suitable to the
situation. For instance, if background noise or other disturbances make their way into the recording, filtering them out should be desirable; perhaps by using an autocorrelation algorithm or a simple high-pass filter, depending on the situation. Using a third metric is optional. If one is to be used, it could either be based on the same audio recording or else some other recording (such as the EGG used in this study), which should be synchronised with the audio recording.

2.4 Electroglottography

Electroglottography (EGG) measures the conductivity across the larynx, the part of the throat where the vocal chords are situated [7, 11–15]. One electrode is placed on the skin on either side of the Adam’s apple (cf. Fig. 6). A low and rapidly changing voltage is applied, creating an AC current with a frequency in the MHz range. This allows for sampling the conductivity across the larynx several hundreds or thousands of times per phonation cycle. The main reason for the conductivity varying is the opening and closing of the vocal folds, resulting in varying contact surface area and thus conductivity, during each phonation cycle [12].

![Figure 4: Example signals. Top: EGG signal. High values correspond to high conductivity (large vocal fold contact area). Bottom: Derivative of EGG (DEGG). Sharp peaks correspond to the folds rapidly closing, simultaneously along their full length; opening is slower and more zipper-like, from front to back.](image)

Some characteristics of phonation can be seen directly in the EGG signal. For instance, the open phase with minimum contact between the vocal folds is visible as a stretch of minimum signal; this could be entirely flat and is then essentially a clipping phenomenon where the EGG signal makes no distinction between different degrees of opening as long
as the vocal folds have reached minimum contact area [11, 14]. In other cases, this portion is more rounded meaning some small changes in contact area take place at all times (see Fig. 4). It is theoretically possible to calculate the contact quotient during a cycle from this information, but attempts have shown that the results depend more than desirably on the exact algorithm used [14].

The exact shape of the vocal fold vibrations cannot be measured using EGG but different modes of vibration have been shown to produce distinct patterns in the EGG signal [10]. Other alternatives than EGG have also been explored in the context of teaching singing techniques, such as surface contact electromyography. This technique gives more information about the muscle activations in the throat than the EGG, which can be used to detect the larynx moving up and down but does not say which muscles cause the movements [16].

One important aspect of using the EGG signal to analyse vocal fold vibrations is the necessary pre-processing of the raw signal [10, 15]. First of all, the variable conductivity resulting from the vocal folds’ varying contact surface causes a relatively small ripple in the carrying current. It is this small variation that is of interest, and it must be extracted from the raw signal. This is so integral to any EGG measurement that it is nowadays built into the hardware. The next step consists of dividing the extracted signal into separate cycles. The repeating waveform of the EGG signal corresponds to vibration cycles for the vocal folds and it is important that the cycle lengths in the EGG signal match those of the actual vibrations; the definition of the beginning of a cycle is less important [15]. One of the most easily identifiable events of a cycle is the sudden increase in conductivity as the vocal folds make full contact [11]. This event is sometimes thought of as the beginning of a cycle and adjacent cycles are usually similar in this part.

2.5 The FonaDyn Software

The present study builds upon a previous work that aimed to develop a software tool capable of analysing EGG signals and providing real-time feedback to the singer [15]. The tool, called FonaDyn (see Fig. 5), is written in the SuperCollider programming language, which is especially developed for real time audio applications.

FonaDyn can operate both with actual real-time input and with pre-recorded files played back. In this study, most of the work was conducted offline, with pre-recorded files. The standard input is then an audio file (.wav format) with two channels; the first channel contains the audio and the other channel contains the EGG data having been simultaneously recorded.
The audio signal is used in an autocorrelation algorithm computing a clarity metric, which dictates which parts of the file contain legitimate phonation and which parts are too turbulent to be useful (typically background noise, coughs etc.). The threshold for acceptable clarity is set in the source code. It can thus be adapted to different levels of acceptance, if the recording environment and/or hardware result in noise levels that match badly with the previous setting.

Figure 5: FONADYN GUI (dividing borders added). Top: Data import/export and cycle detection controls. Middle left: not used here. Centre: EGG cycle display. Bottom left: Fourier Descriptor display (phase plotted vs. amplitude) and cluster member counts (bars). Bottom right: VRP being constructed (various displays available; see Fig. 3).

The EGG signal is first split into separate cycles using one of two available algorithms. For this study, the double-sided peak-tracking algorithm was used. With this algorithm, the signal is derived once (see Fig. 4) since this should improve the cycle detection precision and reliability [13]. By demanding maximum and minimum derivative peaks to alternate, more eccentric EGG waveforms with multiple local maxima during each cycle can still be correctly split into cycles. However, there are still waveforms with unusual shapes that are not correctly classified by this algorithm. Here, the correctness of the cycle separation for the files in question was not examined since the idea was to mimic a live setting, where no such inspection is possible.

Next, each EGG cycle is subject to discrete Fourier transform and decomposed into Fourier Descriptors. These descriptors are the magnitudes and phases of each Fourier component of the EGG cycle. A cycle composed of 5 sinusoidal partials will thus be
expressed as a vector with 10 dimensions (5 amplitudes and 5 phases). When decomposing a cycle into Fourier components, the magnitudes of higher partials are much smaller than for the first few partials, meaning that an EGG cycle can be rather well represented by only the first few partials (the user chooses the number of partials to be kept). The total power of the signal is also computed. From this value, the power of each kept partial is then subtracted, leaving the power contained in all the higher partials that were not kept. This number makes up a final descriptor, representing the smoothness of the cycle and the amount of information not captured by the other Fourier Descriptors.

The incoming Fourier-decomposed EGG cycles accumulate as points in a vector space; the user choosing to keep \( n \) partials results in a \((2n)\)-dimensional space. The idea is that characteristic vibratory states of the vocal folds result in characteristic EGG waveforms, decomposing into characteristic Fourier Descriptors, so that the \((2n)\)-dimensional points produced tend to group together with others produced by the same characteristic vibratory state of the vocal folds. Over time, clusters should begin to crystallise in the vector space, indicating how many characteristic vibratory states there were to begin with and what they typically look like. Note that these states should be independent of the sung pitch: all calculations are done after normalising the EGG signal and separating it into cycles, and all the Fourier Descriptors are expressed relative to the fundamental partial (which is always assigned amplitude 1 and phase 0). This means the EGG cycles will look essentially the same when the singer alters the pitch a semitone, if the vibratory pattern remains qualitatively the same. As a counterexample, when switching to a falsetto vibration of the vocal folds, even when retaining the same pitch, the vibratory pattern changes greatly [10].

Since FONADYN is made to operate in real time, the classification of the accumulating clusters is continuously updated as more points become available. The algorithm used to assign incoming points to clusters is a SUPERCOLLIDER built-in implementation of \( K \)-means. This algorithm keeps track of the \( K \) currently assigned clusters’ members and mean values, finds the cluster most suitable for a new point, assigns it as a member of that cluster and updates the mean value and member count of the cluster. The algorithm can thus be pointed in the wrong direction if the first few samples are unrepresentative of the entire corpus. However, taking all points into consideration is impossible until the entire recording is over, rendering real-time feedback impossible. This greater influence of early samples is thus unavoidable, but can also be used for good by seeding the clustering with suitable examples or predictions.
3 Method

First, relevant features of CCM singing were identified through discussion and preliminary measurements. Next, audio and EGG recordings were made of subjects demonstrating the chosen features. Finally, the data was treated and evaluated to see if the chosen features were indeed detectable in the EGG signal.

3.1 Identifying Relevant Features

This project aimed to bring EGG research and technology closer together with singing practitioners and teachers. The vocal fold vibrations being studied should thus be both detectable by EGG measurements and relevant to practitioners. This was achieved by experts from both fields discussing and agreeing upon a set of phonation candidates.

3.1.1 Feature Candidates

It was decided early on that the project should focus on unhealthy phonation; more specifically, on ways of singing that should be avoided in order to retain vocal health. The focus was thus on preventing voice damage. Among the many candidates, it was generally not known whether the error was a) primarily evident in the vibration pattern of the vocal folds, b) indirectly affecting the vibration pattern, or c) without noticeable effect on the vocal fold vibration pattern. Among the suggested detrimental singing techniques were excessive tongue root tension, bad balance of thyroarytenoid or cricothyroid muscle activation and high levels of pressedness or breathiness. Before selecting a feature to be in focus for the remainder of the study, it was necessary to know which of these, if any, had a detectable influence on the EGG signal.

3.1.2 Selection Process

Several recording sessions were spent with experts from both involved fields, recording samples of a single subject performing the suggested detrimental singing mistakes and establishing experimental protocols and setup. The resulting audio and EGG recordings were discussed continuously during sessions and saved for further analysis. It became evident that some of the suggested features were less suitable for further study. Tongue root tension had no readily detectable effect on the EGG waveform. Leaky phonation added a lot of noise and gave an overall weak signal. It was decided that the project would focus on the differences between pressed and normal phonation, which has been recently investigated in speech using EGG [17]. For ethical reasons it was also clear that any subject participating in the study to record examples of detrimental singing technique
should do so only for very short durations, in order to avoid the damage the techniques eventually incur.

3.2 Recording Setup and Protocol

The main data collection for the project was done during two separate sessions. The first session was held with four female subjects of American and Northern European origin and varying ages. All subjects had had previous training in singing and singing pedagogy; the purpose of this session was to collect data allowing for comparisons between subjects. The second session was held three months after the first and with one of the original four subjects. The purpose of this session was to collect data allowing comparisons of the same subject over time. Before the recording began during the first session, the subjects held a joint warm-up session while discussing their respective interpretations of the terms \textit{pressed} and \textit{normal} voice quality. This was done in an attempt to clarify the definitions and achieve similar performances from the different subjects. For the second session, the warm-up did not include any discussion of the phonation types, since the purpose of the session was to explore if a subject would remain consistent over time or not.

3.2.1 Hardware Setup and Calibration

During all recordings, the subject was standing upright and free to assume a comfortable posture. For sound pressure level calibration purposes, the CM3 cardioid condenser microphone was mounted on a stand in front of the subject, 30 cm from their mouth. A second stand was positioned so as to touch the back of the subject when the correct 30 cm distance was maintained (see Fig. 6). It was possible to approach the microphone (although the subjects were instructed not to), possibly affecting the SPL readings during later parts of recordings, but never the EGG signal. The system was calibrated using an AZ8922 SPL meter held next to the microphone as a sinusoid tone at 440 Hz was played through a loudspeaker some distance away. The microphone amplification was adjusted until the FONADYN cursor hovered at a level matching that of the SPL meter (C weighted).

An RME Fireface UCX soundcard was used to capture both the audio from the microphone and the EGG signal from a Glottal Enterprises EG2 electroglottograph on separate channels. The EGG signal was also sent to a Tektronix TDS210 oscilloscope. For each subject, placement of the EGG electrodes was adjusted until a satisfactorily clear waveform was visible on the oscilloscope display and the EGG signal displays indicated centred vertical placement and sufficient signal strength without warning for clipping. Contact gel was used to improve skin conductivity.
Figure 6: Recording setup. EGG electrodes (middle) were fastened on both sides of the larynx using a Velcro strap. Two microphone stands were positioned so that when the back of the subject touched one of them (left) the other (right) held the recording microphone 30 cm away from the mouth of the subject.

A Roland PC-200 keyboard and JV-1010 sound module were used to generate reference pitches; the subject was free to use the keyboard and the resulting sound was sent to a pair of headphones worn by the subject. Optionally, the sound captured by the microphone could be given a digital reverberation and added to the headphone output for subjects who preferred that feedback to the low level of reverb from the dampened walls of the studio. The subject was positioned in front of a secondary computer screen, shielded to show only select portions of the FONADYN GUI (cf. Fig. 5). The actual data collecting laptop computer was turned as to not be visible to the subject.

3.2.2 Software Settings

For the duration of the project, FONADYN v 1.08 was used. During both sessions, the option save recording was toggled to on in order to retain the data for later analysis. Clustering was also active for monitoring purposes. Different cluster parameters were used but none of the resulting cluster sets was saved. Only two-channel audio/EGG files were recorded during the sessions; the recording of such files is in no way affected by the clustering being active or the parameters used. The clustering process in progress was not visible to the subjects but monitored by the researchers for signs of unforeseen problems or possibilities.

During the project, a bug was discovered in the version of FONADYN used during the first session, resulting in all recordings containing a filtered version of the EGG signal instead of the raw values. The filter in question was a 10 kHz low-pass filter meant to reduce noise in an early stage of the cycle detection process. It was judged that this did not alter the phonation cycles in an appreciable way and the data was used. The bug has been fixed and did not affect the second session. The adjustable clarity threshold was kept at 0.98 throughout the project.
3.2.3 Recording Protocol

The subject was equipped with EGG electrodes, held in place by a Velcro strap and coated with a thin layer of conductive gel for improved signal strength. They also wore headphones. Before recording began, each subject was given some time to get accustomed to wearing this rig and interpreting the FONADYN display of their VRP being constructed. Each subject was first instructed to sing an /a/ at some pitch within their chest voice range and to make the cursor hover at around 80 dB SPL. For the actual recording, the subjects performed glissandi, starting from the previously tested pitch, increasing by two semitones, returning to the original pitch, decreasing two semitones further and finally returning to the original pitch. They thus varied the sung frequency continuously approximately 12% both above and below their chosen starting pitch. Instructions were to keep the SPL as steady as possible. This was done using a normal voice quality. Next, the same phrase was repeated but with pressed voice quality. Instructions were to keep pitch and SPL as close as possible to the previous phrase, using normal voice quality. The phrases were recorded in the same file, so that the subject could aim at tracing the same area of the VRP display that they had previously covered using normal voice quality. This was done in order to avoid introducing variations (other than voice quality) between phrases. However, the VRP display was set to not show the resulting cluster so as not to influence the subject. Each type of voice quality was recorded in a minimum of three repetitions of the sung phrase. The subjects were free to either alternate between the two voice qualities or to record all repetitions of normal voice quality first before moving on to record all repetitions of pressed voice quality.

During the first session, after clearing the VRP display, the same procedure was repeated to record a new file. The new recording was done in much the same way but centred on a higher pitch. This was done in order to record all subjects once in what they themselves felt to be chest and head phonation. In this way, possible differences between registers like those seen in [10] could be studied. Four test subjects were thus recorded in 8 separate files, containing in total slightly over 24 phrases, each containing several thousands of cycles. During the second session, several chest voice samples were recorded for the same subject (see below). This built up fatigue before any high-pitched recordings were made. In order to avoid voice damage, the session ended; no head voice recordings were made.

To help each subject be more consistent in the voice quality alterations during the first session, all four subjects were present during all recordings and allowed to comment between phrases to ensure that they all approved of the voice qualities being produced. At the beginning of the second session, the single subject had had no chance to discuss the different voice qualities or to hear recordings of them since the previous session.
three months prior. This was done in order to investigate if the voice qualities would remain similar over time and without relying on the consensus of several subjects. Three recordings were made during the second session. First, as stated, with no reminder of the results of the first session. Second, with the subject having listened to a recording of herself performing both voice qualities from the first session. Finally, receiving real-time visual feedback based on clusters trained on recordings of herself during the first session. This also allowed for discussing the overall impression of the real-time feedback, of how well it agreed with internal sensory feedback and how useful it could potentially be as a tool for teaching.

3.3 Data Processing

The data processing was divided into three main stages. First, the recorded two-channel audio/EGG files were trimmed of excess content and used to construct training and evaluation material. Next, the training material files were used in FONAdyn to produce cluster sets. Finally, the evaluation material was used as substitutes for live input and classified as normal or pressed according to the different cluster sets (see Fig. 7).

![Diagram of data processing](image)

**Figure 7: Stages of data processing.** Each subject recorded one high-pitched and one low-pitched file, used to construct training and evaluation files, sent to FONAdyn to first make clusters and then label cycles as either normal or pressed. Both clusters and evaluated cycles were further analysed in MATLAB.
### 3.3.1 Constructing Training and Evaluation Material

As previously stated, Fonadyn is capable of real-time feedback without pre-training. Each cycle of incoming data is evaluated according to clusters; these clusters can be trained from scratch, starting from the first cycles of incoming data and updating continuously as long as more data is being input. However, controlling this process to get good clusters can be difficult. Another option is to import ready-made clusters that have been previously trained on some recorded file that should be representative of the type of singing to be evaluated. These pre-made clusters could be based on recordings of the singer to be given feedback, or perhaps clusters could be made that work for any singer. If this is possible or not depends on how consistent phonation types are between singers. This was explored here by separating the training and evaluation steps to simulate a situation where feedback is given using clusters trained in advance. The recorded files were thus used to construct training files (to be used for making the clusters) and evaluation files (to be used for simulating the input during a session where feedback is expected).

The number of repetitions of each phrase and their exact order differed somewhat between subjects. To standardize the files, the minimum number of three repetitions was kept for each voice quality and parts containing talk and background noise were removed. Left were three repetitions of the phrase using normal quality \(N_1, N_2\) and \(N_3\) and three repetitions of the same phrase using pressed quality \(P_1, P_2\) and \(P_3\).

A Jack-knife approach was used to ensure that no phrase was evaluated using a cluster set obtained by training on that same phrase. The training and evaluation sets were constructed in pairs; each pair of normal and pressed phrases was used to make separate evaluation files, and the remaining two normal and pressed phrases were used to construct the corresponding training file. For instance, the files containing only \(N_1\) and \(P_1\), respectively, were evaluated using clusters trained on the file \(N_2P_2N_3P_3\). The voice quality phrases in the training files were always alternated; this was done to introduce both qualities relatively early in the file (by presenting examples of all expected variations early, the \(K\)-means clustering should perform better than if one type is first cemented for many cycles before seeing other types). Finally, all six normal and pressed phrases were made into a single training file. The clusters obtained from such a file were used for evaluating phrases from other subjects, in an attempt to compare the voice quality differences between subjects. Each subject thus gave rise to four training files \((N_1P_1N_2P_2, N_2P_2N_3P_3, N_3P_3N_1P_1\) and \(N_1P_1N_2P_2N_3P_3)\) and six evaluation files \((N_1, N_2, N_3, P_1, P_2\) and \(P_3\)), to be used in appropriate combinations so that training and evaluation files have no overlap. An example of an original \(N_1N_2N_3P_1P_2P_3\) file and final trimmed and rearranged \(N_1P_1N_2P_2N_3P_3\) file can be seen in Fig. 8.
For each new sung phrase, phonation is re-initialised. During phonation initialisation, EGG cycles will gradually establish a steady wave shape but the first few cycles can vary greatly in shape. At the very first phrase, the clusters are not yet well-defined and each such cycle type could end up being assigned as the centroid of their own cluster; they are probably too different to be clustered together by the algorithm. This sets the clustering up in a bad way. When the cycle shapes have stabilised, the cycles of actual interest will vary when the voice quality changes, but often far less than the first erratic cycles did. This can cause all steady-state cycles to end up in a single cluster instead of being correctly separated.

To remedy this, FONA Dyn lets the user manually restart the clustering at any time. By resetting the clustering when the cycle shapes have already stabilised, more of the resulting cluster centroids will be small variations of steady phonation cycles, enabling the voice quality variations of interest to fall into separate clusters; fewer clusters will consist of very erratic cycles (garbage). More garbage cycles will be generated at the end of the phrase (and at the beginning of the next phrase) - by then, however, many steady-state cycles will have been seen by the clustering algorithm. The garbage cycles will then perturb the cluster centroids only slightly. Some surplus clusters can be used as garbage collectors; the number of clusters used should always exceed the number of expected phonation types for this very reason.

In this study, the need for repeatability discouraged the use of such manual resets. However, this action is meant to be used by a pedagogue and so it should be at least simulated here. Thus, the beginning of each training file was cut to start within the first phrase, skipping the unsteady phonation initialisation phase (cf. Fig. 8).

3.3.2 From Training Material to Clusters

The end goal was to have FONA Dyn give singers feedback telling them when their singing was pressed enough to warrant caution. The training material was taken from the first session and consisted of recordings of singers instructed to sing either in a normal way...
or in a way that was too pressed. No formal criteria were given for what constituted pressedness; each recorded phrase was labelled as normal or pressing entirely based on the intention of the subject singing. It was unknown whether the cycles in these phrases could be successfully separated into distinct clusters. Failure to separate cycles into intended clusters would indicate either that the software could not classify cycles with sufficient precision, or that the singers’ perceived and intended differences between normal and pressed voice quality did not correspond to qualitative differences in the EGG cycles. A successful separation would indicate a qualitative difference in the Fourier Descriptors between different intended voice qualities. These differences could then be similar or different across subjects; if different, cluster sets would be singer specific and each student would need their own tailor-made set; if similar, cluster sets could perhaps be reused between students or even combined into a general library for voice evaluation.

Training and evaluation can be done simultaneously in FONADYN by always using the $n - 1$ already observed cycles as training material to evaluate the $n$:th cycle, but there are some limitations to this. First, the clusters usually get more well-defined over time as more cycles are seen. Conversely, early on in simultaneous training and evaluation the results are often less accurate. If training is done first using one file, another file can then be evaluated with clusters that are already fully developed and the instability of the early stages affect the evaluation results less. Second, the clustering is done according to user settings. The results vary greatly depending on these settings. The user must find a good number of overtones to use, suiting the cycle differences being looked for. Also the number of clusters to use must be chosen. Using too few clusters, cycles that should ideally be separated into different clusters end up in the same cluster. Using too many clusters, the cycle types to be separated can be split further, into more clusters than desired. The number of clusters must be tuned so that the appropriate level of discrimination is achieved, which varies between subjects, phonation types and amount of garbage cycles. Getting the settings right usually takes several attempts. In simultaneous training and evaluation, many useless evaluation files would be generated while fine-tuning the settings. For the purposes of this study, suitable settings were found for each training file by trial and error, without simultaneous evaluation. The resulting cluster sets were exported for later use in the evaluation step and for direct analysis in MATLAB R2015b.

In order to make the cluster sets more comparable within and between subjects, the number of harmonics used should ideally be the same for every training file. However, no single number of harmonics was found suitable for all files. Using seven harmonics worked well for the recordings of the low-pitched phrases but not for the high-pitched ones. Similarly, four harmonics worked well for all high-pitched recordings but not for the
low-pitched ones. Since no single number was found, different settings were used for the low- and high-pitched recordings: seven and four harmonics, respectively.

Even with common settings for the number of harmonics, the best number of clusters to use will differ between subjects and even phrases. More garbage cycles or finer differences between the relevant cycle types mean more clusters must be added before the appropriate level of resolution is achieved. The relevant clusters are usually few, containing most of the cycles. The remaining clusters act as garbage collectors during training and have no relevance during the evaluation step. In this study, each phrase included several thousand cycles. After exporting each cluster set, any cluster with fewer than 1,000 cycles was removed. Clusters were also sorted according to corresponding cycle type; pressed first, then normal. This was done to simplify the evaluation step. By removing small and ambiguous clusters, each cycle from the evaluation files is clearly labelled as either normal or pressed. Cycles are labelled by cluster index. By sorting (grouping) the clusters first, low indices mean pressed and high indices mean normal, which simplifies the analysis.

3.3.3 From Clusters to Evaluated Cycles

After generating and modifying the cluster sets, each set was loaded into FONADYN together with its corresponding evaluation file. Between 2 and 20 clusters can be used when training a cluster set. Some training files had required the full 20 clusters to be used. However, having removed clusters with fewer than 1,000 members, only a few relevant clusters remained in each set: 2-3 clusters remained for sets trained on the low-pitched files and 4 clusters for the sets trained on the high-pitched files. When importing cluster sets in this manner, the clustering algorithm can be allowed to keep learning (see [15] for details). This option was not used here.

Each cluster set had been trained on both normal and pressed phrases, all from the same subject - one example of a training file would be $N_1P_1N_2P_2$. Each evaluation file consisted of a single phrase from the same subject, intended to be either normal or pressed: this would be $N_3$ and $P_3$ in the same example as before. An ideal evaluation would label each cycle of $N_3$ as normal and each cycle of $P_3$ as pressed. If the evaluation is successful, it indicates that the difference between normal and pressed quality has been consistent between phrases.

For each subject, a cluster set had also been trained based on all six phrases - this would be $N_1P_1N_2P_2N_3P_3$ using the same example as above. This cluster set was used to evaluate phrases recorded by the other subjects. If this evaluation is successful, it indicates that the difference between normal and pressed quality is similar between subjects.
The result of each evaluation was exported and analysed in MATLAB. The labelling of a cycle either agreed with the intention of the singer or contradicted it. Since the numbers of intended pressed and normal cycles were similar in all files, approximately half should be pressed and half should be normal. Randomly labelling each cycle as either pressed or normal should thus result in approximately 50% correctly classified cycles; as would assigning all cycles as pressed (or normal). Only an evaluation clearly exceeding 50% accuracy could be considered successful.

3.3.4 Analysis of Clusters as Vectors

As described in Section 3.3.1, each raw audio/EGG file was used to construct four different training files. From each such file at least two clusters were made. A cluster is defined by its centroid; each centroid is effectively a vector and the vector elements are the Fourier Descriptors (cf. Section 2.5). Treating the cluster centroids as vectors is practical when comparing clusters from different training files or from different subjects. For these comparisons, one pressed quality and one normal quality cluster was included from each training file. This means eight clusters were used from each subject - four normal and four pressed. The clusters were analysed to evaluate how consistent each subject was when alternating between phrases intended to contain either normal or pressed phonation qualities, and how similar cluster behaviour was between subjects.

If a subject has been consistent between phrases, the four normal clusters are similar, the four pressed clusters are similar, and the difference between normal and pressed is similar between the four training files. An attempt was made to quantify these similarities. For points in a vector space, Euclidean distance is a common way to measure distances. We assume that the Fourier Descriptors are independent and use the generalized Pythagorean theorem to calculate distances between cluster centroids. If the two phonation qualities correspond to well-defined EGG waveforms that remain the same between phrases, all normal quality centroids should lie in the same region, all pressed quality centroids should lie in the same region and these two regions should be separate. This translates to small distances between centroids of the same type and larger distances between centroids of opposite types.

If the subjects are confirmed to perform consistently between phrases, the next question is if the changes made from normal to pressed quality are similar between subjects. We describe the change a subject makes between normal and pressed quality using the difference between the normal centroid and the pressed centroid. This difference is a new vector, pointing from normal to pressed. This vector thus defines the direction of increasing pressedness for a particular training file. The four pairs of normal and pressed
centroids from each subject will result in four directions of increasing pressedness per subject; in total, sixteen vectors. If these vectors point in the same direction, this indicates that the EGG signal changes in similar ways between subjects. If they point in different directions, EGG signals change differently for different subjects, meaning clusters for feedback might need to be tailor-made for each singer. As a measure of parallelity we use the length of the projection of one normalised vector onto another. This gives a number between plus one (completely parallel vectors) and minus one (antiparallel vectors).

Both of these comparisons are made of the full centroids. Since each centroid consists of several Fourier Descriptors, it is possible that any differences between normal and pressed quality phonation manifest more in certain Fourier Descriptors. Therefore, apart from the overall positions of the normal and pressed centroids, we also inspect and compare their individual Fourier Descriptors.

4 Results

As previously mentioned, no single number of harmonics could be successfully used for all training files. Therefore, high-pitched and low-pitched recordings were treated separately. Also, the number of clusters used varied between training files. It was not always possible to find a number of clusters such that normal and pressed quality cycles ended up in exactly one cluster each. In order to achieve satisfactory separation, the number of clusters sometimes had to be increased until one cycle type split into more than one cluster before the other cycle type ended up in its own cluster. This meant that even after removing clusters with fewer than 1,000 cycles, some training files generated cluster sets with more than two clusters (interestingly, cycles from pressed quality tended to end up in a single cluster, whereas cycles from normal quality were the ones who split into two or more clusters). As previously mentioned, the vector analysis of cluster centroids used exactly one pressed and one normal quality cluster from each training file. The selection was based on how well normal and pressed quality centroids grouped together. This further selection was not done for the clusters used in FONADYN during the evaluation step.

4.1 Analysis of Clusters as Vectors

4.1.1 Distances between Cluster Centroids

Pairwise Euclidean distances between cluster centroids for each of the four subjects are presented in Figures 9 and 10 for the high-pitched and low-pitched recordings, respectively. The greyscale is normalised in each grid so that black represents the shortest distance and white represents the greatest distance between centroids for that subject.
The actual values corresponding to the shades are shown by the bar on the right hand side of each grid. All grids are symmetrical along one diagonal (the black one - each centroid is at distance zero from itself). The four pressed quality centroids are on the left and top halves of the two axes, respectively. The values in the four grids span different ranges, but it is clear that there are four blocks in each grid; two are lighter and two are darker. In the top left block are the pairwise distances between the four pressed quality centroids; in the bottom right block are the pairwise distances between the four normal quality centroids. The upper right and lower left blocks contain the pairwise distances between cluster centroids of different qualities; one normal and one pressed. Since darker colour represents shorter distance, it is evident that the normal quality centroids are closer to each other than to the pressed quality centroids, and vice versa. In summary, the absolute positions of the normal and pressed quality centroids remain fairly consistent between phrases and so do the distances between normal and pressed centroids.

Figure 9: Pairwise Euclidean distances for cluster centroids derived from the high-pitched training files. Pressed quality centroids are on the left and top halves of the two axes, respectively; centroids of the same kind clearly lie closer together.

Figure 10: Pairwise Euclidean distances for cluster centroids derived from the low-pitched training files. Pressed quality centroids are on the left and top halves of the two axes, respectively; centroids of the same kind clearly lie closer together.

4.1.2 Directions of Increasing Pressedness

Pairwise parallelity measures for vectors pointing towards increasing pressedness are shown in Figure 11 for both high-pitched recordings (left) and low-pitched recordings (right). The greyscale is normalised in each grid so that black represents the worst parallelity and white represents complete parallelity. The actual values corresponding to the shades are shown by the bar on the right hand side of each grid. Both grids are symmetrical along one diagonal (the white one - each vector is completely parallel to itself).
The grids are sixteen cells wide and high, since the comparison is done with vectors from four training files per subjects with four subjects in total. From the left (and top), the first four ticks correspond to vectors from the four training files of the first subject. The next four ticks correspond to vectors from the second subject, and so on. There is a pattern emerging with blocks four cells high and wide that are either mostly white, grey or black. This is because all four vectors from one subject are all approximately equally parallel to all vectors of another subject. The blocks along the diagonal are almost white, meaning the four vectors from each subject are close to fully parallel. Subjects whose vectors are close to parallel (but not quite as close as within a single subject) give rise to grey blocks. Very dark blocks show what subjects have the most different directions of increasing pressedness. In both grids, all the darkest blocks involve the second subject. Note however how the values of the lowest parallellity measures differ between the two grids; one is close to minus one; the other closer to plus one half. The disagreement between the second subject and the remaining subjects is thus much greater in the grid on the left. The measure of −0.8606 corresponds to an angle of 149 degrees, meaning that this subject has a direction of increasing pressedness almost opposite to those of the remaining subjects. In the grid on the right, the corresponding angle is 62 degrees. In this case, it is possible to find a direction more parallel than antiparallel to all vectors from the low-pitched recordings, which might not be true for the high-pitched case because of the larger angles.

Figure 11: Pairwise parallelity measures for directions of increasing pressedness from high-pitched recordings (left) and low-pitched recordings (right). Four values each from subjects 1-4 are ordered from left to right and from top to bottom of the two axes; subject 2 shows the poorest match with the others in both grids, especially in the left grid.
4.1.3 Comparison of Fourier Descriptors

As described in Section 2.5, a cluster centroid consists of the Fourier components of the mean cycle shape for that cluster. For the high-pitched training files, a setting of four harmonics was used. The first of these harmonics is the fundamental frequency and used purely as a point of reference for constructing the Fourier Descriptors, which are amplitudes and phases relative to those of the fundamental. The first harmonic thus has amplitude one and phase zero for all cluster centroids and the Fourier Descriptors of the remaining three harmonics are what set one centroid apart from another. Two of these three overtones are actual overtones with both amplitude and phase; the third has amplitude only and collects the energy not included in the lower overtones. As above, four normal quality and four pressed quality centroids were used from each subject, corresponding to the four training files constructed. The averages and standard deviations for the components of these centroids are shown in Figure 12.

Figure 12: Mean values and standard deviations for Fourier Descriptors of pressed and normal centroids from the high-pitched training files. For all four subjects, indices 1 and 2 are relative amplitudes of the 1st and 2nd overtones; indices 4 and 5 are relative phases for the same overtones; index 3 is the measure of higher-overtone energy.
For the low-pitched training files, a setting of seven harmonics was used. As the fundamental is by definition identical between centroids, the Fourier Descriptors of the remaining six harmonics are what set one centroid apart from another. Again, the last one is not a proper overtone - it has amplitude only and collects the energy left in all higher overtones. As before, four normal quality and four pressed quality centroids were used from each subject, corresponding to the four training files constructed. The averages and standard deviations for the components of these centroids are shown in Figure 13.

Figure 13: Mean values and standard deviations for Fourier Descriptors of pressed and normal centroids from the low-pitched training files. For all four subjects, indices 1-5 are relative amplitudes of the first five overtones; indices 7-11 are relative phases for the same overtones; index 6 is the measure of higher-overtone energy.

### 4.2 Accuracy of Cluster-based Classification

Each cluster set was used to evaluate a number of phrases, each intended to consist entirely of normal quality cycles or entirely of pressed quality cycles. Below follow tables containing accuracies for normal phrases, for pressed phrases, and total accuracy for both pressed and normal phrases for each cluster set. The training and evaluation files were
always both from high-pitched recordings or both from low-pitched recordings, because of the different number of harmonics apparently needed to successfully treat the different registers. The tables in Sections 4.2.1 and 4.2.2 below present results for cluster sets trained on subjects 1-4, in order. All tables follow the same structure. From left to right, their columns contain: What subject was used for constructing the evaluation files; the total number of intended pressed quality cycles in those files; the number and percentage of correctly identified pressed cycles; the total number of intended normal quality cycles; the number and percentage of correctly identified normal cycles; total accuracy for all cycles, pressed and normal, as a percentage.
### 4.2.1 High-pitched Recordings

Table 1: Evaluation results: clusters trained on high-pitched files from subject 1.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15690</td>
<td>9788 (62%)</td>
<td>14818</td>
<td>10972 (74%)</td>
<td>68%</td>
</tr>
<tr>
<td>2</td>
<td>12590</td>
<td>9263 (74%)</td>
<td>12859</td>
<td>6699 (52%)</td>
<td>63%</td>
</tr>
<tr>
<td>3</td>
<td>16799</td>
<td>15062 (90%)</td>
<td>17602</td>
<td>15868 (90%)</td>
<td>90%</td>
</tr>
<tr>
<td>4</td>
<td>13015</td>
<td>8971 (69%)</td>
<td>12378</td>
<td>12195 (99%)</td>
<td>83%</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results: clusters trained on high-pitched files from subject 2.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15690</td>
<td>12269 (78%)</td>
<td>14818</td>
<td>8288 (56%)</td>
<td>67%</td>
</tr>
<tr>
<td>2</td>
<td>12590</td>
<td>6866 (55%)</td>
<td>12859</td>
<td>12468 (97%)</td>
<td>76%</td>
</tr>
<tr>
<td>3</td>
<td>16799</td>
<td>15246 (91%)</td>
<td>17602</td>
<td>183 (1%)</td>
<td>45%</td>
</tr>
<tr>
<td>4</td>
<td>13015</td>
<td>13008 (100%)</td>
<td>12378</td>
<td>130 (1%)</td>
<td>52%</td>
</tr>
</tbody>
</table>

Table 3: Evaluation results: clusters trained on high-pitched files from subject 3.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15690</td>
<td>491 (3%)</td>
<td>14818</td>
<td>14790 (100%)</td>
<td>50%</td>
</tr>
<tr>
<td>2</td>
<td>12590</td>
<td>743 (6%)</td>
<td>12859</td>
<td>11981 (93%)</td>
<td>50%</td>
</tr>
<tr>
<td>3</td>
<td>16799</td>
<td>9719 (58%)</td>
<td>17602</td>
<td>17133 (97%)</td>
<td>78%</td>
</tr>
<tr>
<td>4</td>
<td>13015</td>
<td>1148 (9%)</td>
<td>12378</td>
<td>12328 (100%)</td>
<td>53%</td>
</tr>
</tbody>
</table>

Table 4: Evaluation results: clusters trained on high-pitched files from subject 4.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15690</td>
<td>15137 (96%)</td>
<td>14818</td>
<td>3837 (26%)</td>
<td>62%</td>
</tr>
<tr>
<td>2</td>
<td>12590</td>
<td>3114 (25%)</td>
<td>12859</td>
<td>11445 (89%)</td>
<td>57%</td>
</tr>
<tr>
<td>3</td>
<td>16799</td>
<td>15265 (91%)</td>
<td>17602</td>
<td>5133 (29%)</td>
<td>50%</td>
</tr>
<tr>
<td>4</td>
<td>13015</td>
<td>11692 (90%)</td>
<td>12378</td>
<td>12159 (98%)</td>
<td>94%</td>
</tr>
</tbody>
</table>
### 4.2.2 Low-pitched Recordings

Table 5: Evaluation results: clusters trained on low-pitched files from subject 1.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9445</td>
<td>9408 (100%)</td>
<td>9154</td>
<td>8315 (91%)</td>
<td>95%</td>
</tr>
<tr>
<td>2</td>
<td>6524</td>
<td>70 (1%)</td>
<td>7677</td>
<td>7641 (100%)</td>
<td>54%</td>
</tr>
<tr>
<td>3</td>
<td>11686</td>
<td>6470 (55%)</td>
<td>10036</td>
<td>9842 (98%)</td>
<td>75%</td>
</tr>
<tr>
<td>4</td>
<td>8042</td>
<td>7996 (99%)</td>
<td>7600</td>
<td>5778 (76%)</td>
<td>88%</td>
</tr>
</tbody>
</table>

Table 6: Evaluation results: clusters trained on low-pitched files from subject 2.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9445</td>
<td>9441 (100%)</td>
<td>9154</td>
<td>11 (0%)</td>
<td>51%</td>
</tr>
<tr>
<td>2</td>
<td>6524</td>
<td>6450 (99%)</td>
<td>7677</td>
<td>5727 (75%)</td>
<td>86%</td>
</tr>
<tr>
<td>3</td>
<td>11686</td>
<td>11686 (100%)</td>
<td>10036</td>
<td>140 (1%)</td>
<td>54%</td>
</tr>
<tr>
<td>4</td>
<td>8042</td>
<td>8039 (100%)</td>
<td>7600</td>
<td>0 (0%)</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table 7: Evaluation results: clusters trained on low-pitched files from subject 3.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9445</td>
<td>9440 (100%)</td>
<td>9154</td>
<td>7534 (82%)</td>
<td>91%</td>
</tr>
<tr>
<td>2</td>
<td>6524</td>
<td>124 (2%)</td>
<td>7677</td>
<td>7606 (99%)</td>
<td>54%</td>
</tr>
<tr>
<td>3</td>
<td>11686</td>
<td>8221 (70%)</td>
<td>10036</td>
<td>9828 (98%)</td>
<td>83%</td>
</tr>
<tr>
<td>4</td>
<td>8042</td>
<td>8030 (100%)</td>
<td>7600</td>
<td>3187 (42%)</td>
<td>72%</td>
</tr>
</tbody>
</table>

Table 8: Evaluation results: clusters trained on low-pitched files from subject 4.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number of P-cycles</th>
<th>Found P-cycles</th>
<th>Number of N-cycles</th>
<th>Found N-cycles</th>
<th>Total accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9445</td>
<td>9441 (100%)</td>
<td>9154</td>
<td>2728 (30%)</td>
<td>65%</td>
</tr>
<tr>
<td>2</td>
<td>6524</td>
<td>843 (13%)</td>
<td>7677</td>
<td>7611 (99%)</td>
<td>60%</td>
</tr>
<tr>
<td>3</td>
<td>11686</td>
<td>7115 (61%)</td>
<td>10036</td>
<td>9819 (98%)</td>
<td>78%</td>
</tr>
<tr>
<td>4</td>
<td>8042</td>
<td>7072 (88%)</td>
<td>7600</td>
<td>6924 (91%)</td>
<td>89%</td>
</tr>
</tbody>
</table>
4.2.3 Long-term Consistency

One of the subjects returned for a second recording session, three months after the original session. This time, only low-pitched files were recorded. Having seen that subjects had been fairly consistent between phrases during the first session, the next question was whether or not this consistency persisted over longer periods of time. To see this, the recordings from the second session were used to make new evaluation files. Material from the first session was used to construct five different cluster sets:

1. One pair of normal quality and pressed quality centroids (derived from one training file made during the first session, of the same subject attending the second session).

2. One pair of normal quality and pressed quality centroids (derived from all four first-session training files of the subject present during the second session by averaging the four resulting normal and pressed centroids, respectively).

3. Four pairs of normal and pressed quality centroids (each derived from one training file and different subjects from the first session), added to the same cluster file.

4. Four pairs of normal and pressed quality centroids (each derived from all four training files of the four different subjects, by averaging the clusters of each subject separately), added to the same cluster file.

5. One pair of normal and pressed quality centroids (made by averaging the four pairs of the previous cluster set).

At the start of the second session, the subject was given no reminder of the voice qualities used during the first session other than their names. After warm-up, one set of normal and pressed quality phrases was recorded. Next, the subject was allowed to listen to recordings of herself from the first session, and another set of phrases was recorded. Finally, the VRP display of FONADYN was made visible. The arrangement was such that cycles identified by the algorithm as pressed or normal would colour the cursor red or blue, respectively. After getting used to this feedback for a few minutes, the subject was instructed to record a final set of phrases, making her best effort to keep the cursor blue during the normal phrases and red during the pressed phrases.

These three different attempts were then used as evaluation material for the five different cluster sets described above. The resulting accuracies are presented in Table 9, where the cluster sets have been sorted in order of decreasing accuracy.
Table 9: Correctly identified cycles using five different cluster sets, sorted in order of decreasing accuracy. The subject first had no reference to the first recording session, then was allowed to listen to recordings, then finally given real-time visual feedback.

<table>
<thead>
<tr>
<th>Subject status</th>
<th>Set 2</th>
<th>Set 1</th>
<th>Set 3</th>
<th>Set 4</th>
<th>Set 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No clues</td>
<td>54%</td>
<td>54%</td>
<td>53%</td>
<td>53%</td>
<td>53%</td>
</tr>
<tr>
<td>Listening</td>
<td>75%</td>
<td>69%</td>
<td>57%</td>
<td>54%</td>
<td>53%</td>
</tr>
<tr>
<td>Feedback</td>
<td>94%</td>
<td>93%</td>
<td>90%</td>
<td>87%</td>
<td>84%</td>
</tr>
</tbody>
</table>

5 Discussion

5.1 Normal and Pressed Quality Cycle Shapes

It is interesting that the same number of harmonics proved useful for all high-pitched recordings when constructing cluster sets. Not only did the same number of harmonics work for all files from one subject; four harmonics worked well for high-pitched files from all subjects. Likewise, seven harmonics worked well for all low-pitched files across all subjects. Cycles could in theory be evaluated using clusters constructed using any number of harmonics, but it stands to reason that better results can be hoped for if training and evaluation files separately seem to work well with the same number of harmonics.

The need to use different numbers of harmonics for high and low registers has both theoretical and practical implications. First, it indicates that the hypothetical division of phonation into normal and pressed quality may be too simple and that there are in fact several domains in which the differences between normal and pressed quality manifest differently in the EGG waveform. Second, it becomes difficult to give feedback over the entire VRP since cycles are assigned to their closest cluster, based on the Fourier Descriptors. This is done in a vector space of set dimensionality, dictated by the number of harmonics used. It is unclear how to compare cluster matches if cluster centroids have different dimensionality, as would be the case if combining four-harmonic cluster centroids for high register with seven-harmonic cluster centroids for low register.

Looking at the average values for Fourier Descriptors in Figs. 12 and 13, some trends are hinted at in the differences between normal and pressed cluster centroids. For the high-pitched clusters, the amplitude of the first overtone seems generally higher for the normal quality centroids, as opposed to the low-pitched clusters where the opposite seems to be the case. The low number of descriptors and subjects make it difficult to say more about the high-pitched clusters. For the low-pitched clusters, the descriptor capturing
the remaining high-overtone energy seems to generally have higher values for the pressed quality centroids. However, the greatest and most consistent difference lies in the relative phases, which are all larger for the pressed quality centroids for all subjects. Combined with the larger number of overtones needed to successfully separate pressed and normal quality in the low-pitched case, this tells us something about the cycle shape differences between normal and pressed quality. First, more overtones are needed to describe smaller details in cycle shapes. Second, higher relative phases of overtones are needed for skewing the cycle shape, making it less sinusoidal and more asymmetric. Figure 14 shows examples of typical normal and pressed cycle shapes for high-pitched and low-pitched files. For both high- and low-pitched files, pressed quality cycles are wider, corresponding to a longer closed phase of the cycle, and more asymmetric due to increasing differences between closing and opening phases.

![Figure 14: Examples of typical cycle shapes. Left: High-pitched example of a normal (dashed line) and pressed quality (solid line) cycle. Right: Low-pitched example of a normal (dashed line) and pressed quality (solid line) cycle.](image)

One difference between the examples shown in Figure 14 is the formation of a shoulder in the cycle shape for low-pitched pressed quality cycles. Representing this shape is more readily done using six overtones than with only three, possibly explaining the need for more harmonics when training on the low-pitched files (see Figure 15).

![Figure 15: Reconstructing a stylised EGG cycle (with shoulder) using varying numbers of Fourier components. Left: Three components. Right: Six components. Notice how increasing the number of components improves how well the reconstruction matches the original at the shoulder.](image)
5.2 Cluster Properties and Evaluation Accuracies

The cluster centroids constructed from the training files seem fairly well separated into pressed and normal quality, as indicated by Figures 9 and 10. Subjects seem rather consistent but there is some variation between phrases, which was the reason for investigating the usefulness of cluster sets made by averaging all cluster sets from the same subject; any subsequent recording should be more likely to resemble the average of previous observations than any particular previous observation. Based on Table 9, this seems to hold some merit. The same cannot be said for when using cluster centroids from all subjects, where the averaging process slightly decreases the accuracy rather than improve it. However, using one pair of normal and pressed cluster centroids for each observed subject becomes troublesome when the number of subjects grows. This was the incentive for using all four subjects to construct a single pair of normal and presses cluster centroids, which can be done in the same way no matter the number of subjects. The results were surprisingly good when evaluating the file from when the subject received visual feedback.

Less surprisingly, this file was the most successful for all cluster sets used for evaluation. When no reminder had been given of the phonation qualities, all cluster sets gave equally poor results with accuracies close to 50%, indicating that although the phonation qualities remained consistent between phrases, this need not be true over longer periods of time. Interestingly, after letting the subject listen to a recording of herself from the first session, cluster sets based on her voice alone saw much greater improvements than those based on all subjects. On the other hand, when given visual feedback based on a cluster set (based on her voice alone, in fact set 1 of Table 9), all cluster sets performed much better. This could perhaps be expected for cluster sets containing the exact same pair of normal and pressed quality cluster centroids being used for the feedback (sets 1 and 3 of Table 9) where these clusters could directly absorb all the cycles. Also sets containing the average clusters of the same subject (sets 2 and 4 of Table 9) should lie close to the clusters used for the feedback. However, even the set containing the average of all subjects (set 5 of Table 9) gave an accuracy of 84% despite having only a distant relation to the clusters used for the visual feedback being followed during the recording. This leaves open the possibility that an average pair of normal and pressed quality cluster centroids could still be useful, perhaps increasingly so with increasing numbers of subjects used.

When both training and evaluation files were taken from the same subject, accuracies of 68-94% were achieved for the high-pitched files and 83-95% for the low-pitched files (see Tables 1 - 8). These accuracy levels are rather high, given the many simplifications having been made. For instance, each recorded phrase has been seen as completely comprised of either normal or pressed quality cycles. Pressedness varies during a phrase and with
accumulated fatigue, meaning all files contain cycles of varying kinds. Also, garbage-collecting clusters were removed, meaning even erratic cycles are labelled as normal or pressed. They are neither and introduce a level of noise by randomly being assigned to either category.

When training and evaluation files were taken from different subjects, varying accuracies of 45-91% were achieved. It can be noted in Tables 1 - 8 that low accuracies (around 50%) are associated with virtually all normal cycles being correctly identified and all pressed cycles being incorrectly identified, or vice versa. There are no cases where normal and pressed cycles are both mostly incorrectly identified. This makes sense given the implications of Figure 11; the direction of increasing pressedness is fairly consistent for all subjects, except for the second subject in the high-pitched case. If the direction of increasing pressedness is the same for training and evaluation files, then normal and pressed cycles will never both be mostly incorrectly identified, since that would require the directions of increasing pressedness to be antiparallel. This was almost the case for the second and third subject; the blocks in Figure 11 corresponding to comparing these subjects are the darkest in the high-pitched case. This matchup also resulted in the lowest accuracy of 45% (see Table 2). The antiparallelity could have caused an even lower accuracy, depending on the absolute positions of the cluster centroids involved; a further illustration of the way centroid position and parallelity measure should affect evaluation accuracy is given in Figure 16.

![Figure 16: Left to right: 1) Illustration of both subjects having similar directions of increasing pressedness. 2) Bad overlap, giving accuracy of around 50% regardless of parallelity. 3) Good overlap, giving accuracy of around 90%. 4) Good overlap but antiparallel directions of increasing pressedness, giving accuracy close to 0% (not seen in this project).](image)

Interestingly, as seen in Tables 1 and 7, subjects were sometimes more successfully evaluated based on each other’s training files than based on their own training files. For instance, in the high-pitched case, subject 3 is is evaluated with better accuracy using
training files from subject 1 than either evaluating subject 1 using training files from subject 1 or evaluating subject 3 using training files from subject 3 (cf. Table 1). That the evaluation works better across subjects than within subjects is surprising but possibly coincidental. Regardless, this bodes well for future attempts at finding clusters that work well across subjects.

5.3 Discussion of Sustainability

The requirements for a Degree of Master of Science in Engineering state that the thesis work should demonstrate awareness of social and ethical aspects as well as of how the work connects with society’s goals of economically and ecologically sustainable development. The following discussion concerns those requirements.

Regarding ethical aspects, the main issue for this project has been how to study detrimental behaviour in singers without causing the subjects harm. In order to minimise this risk, measurement sessions were kept short and aborted as soon as a subject felt discomfort or fatigue.

According to the International Federation of the Phonographic Industry, global recorded music sales totalled US $15.0 billion in 2015 [18]. Singers who develop voice disorders must either retire early or go through rehabilitation - this affects the music industry in particular but also society in general since tax revenue goes down while costs for vocal health care increase with increasing incidence of voice damage.

It is also noteworthy that voice disorders concern a wider range of professions than singers alone. In the US, close to 20% of the working population (primarily salespeople and teachers) has been identified as consisting of professional voice users who depend on their voice for their livelihood [19]. Teachers are also reported to suffer approximately twice the risk of developing a voice disorder as compared to the general public [19–21].

Finally, it is difficult to say how this work impacts ecological sustainability. Software development in itself only leaves a small footprint since no large amount of resources or transportation is used. An international gathering of singing pedagogues was taken advantage of in order to collect data from subjects from around the world without needing extra trips between countries.
6 Concluding Remarks

Overall, it was rather easy to make the clustering algorithm recognise even vaguely defined phonation types with accuracies sometimes as high as 95%. The results varied somewhat between subjects; at the same time, some benefits of averaging values were seen. It remains to be seen whether a larger number of subjects will give similar results, and if data from more subjects can be combined to improve results or if the algorithm had better be given clusters tailor-made to each singer.

The visual feedback should ideally become useful in a student-teacher situation by clearly and objectively visualising phonation qualities. Apart from recording training and evaluation files, part of the second session was spent discussing the perceived user friendliness of FONADYN. The feedback corresponded rather well with phonation quality but was severe in its judgments and not so easy to follow for a singer. The current display shows a history VRP where cells are coloured by their most commonly seen phonation quality, with the cursor having the colour of the current phonation quality. The small cursor and coloured background make the feedback less obvious. However, the overall impression was that the tool might indeed become useful with an improved visualisation scheme.

Of course, this is merely a first exploration of the potential of EGG as a means to improve teaching of CCM singing technique while maintaining vocal health. More subjects must be studied, the phonation qualities that can be detected using EGG must be further explored, the type of visual feedback must be improved upon, and the EGG apparatus in itself optimised to interfere as little as possible with the fine adjustments associated with singing.

7 Acknowledgements

I would like to thank professor Sten Ternström, who throughout the project has inspired with his enthusiasm for all things musical and technical, and gone way beyond the formal responsibilities of a supervisor. My thanks also to Ragnhild Bøhn Sørby, who made the project possible through her expertise in CCM and passion for vocal health, to Jeanette LoVetri for her experienced guidance in the recording studio, and to all four singing pedagogues for their patience and participation as test subjects. Finally to Dennis Johansson, creator of FONADYN, to Zimon Mossberg for his crucial feedback during the project, and to the Department of Speech, Music and Hearing where this study was made.
References


[18] International Federation of the Phonographic Industry homepage: http://www.ifpi.org (visited 2016-10-09)


