A time-aligned account of lingual and laryngeal gestures using ultrasound tongue imaging (UTI) and electroglottography (EGG)

Stefano Coretta

May 15, 2017

Contents

1 Introduction

This document is a technical report of the methodology specifically developed for this PhD project. The results of a pilot study testing the methodology will also be discussed. Extracts of this report will be included in the dissertation as part of the "Literature review" and "Methodology" chapters.

2 Purpose

The combination of techniques described in the following sections allows a synchronous mapping of lingual gestures and phonation during speech. Such methodology enables a time-aligned account of the movements of the tongue and the concomitant configuration of the glottis that characterises phonation. These techniques employ ultrasound tongue imaging (UTI) and electroglottography (EGG) for the simultaneous acquisition of articulatory data from, respectively, the oral cavity and the glottis.

2.1 Ultrasound tongue imaging

Ultrasound tongue imaging (UTI) uses ultrasonography for charting the movements of the tongue into a two-dimensional image. In medical sonography, ultrasound waves (sound waves at high frequencies, ranging between 2 and 14 MHz) are emitted from a probe in a fan-like manner, and travel through organic tissue (such as skin and muscles). When the surface of a material with different density is hit, the ultrasound waves are partially reflected, and such "echo" is registered by the probe. These echoes can then be plotted on a two-dimensional graph, where different densities are represented by different shades (higher densities are brighter, while lower densities are darker). The graph, or ultrasound image, will show high density surfaces as very bright lines, surrounded by darker areas (Figure 1). By positioning the ultrasound probe in contact with the submental triangle (the surface below the chin), sagittally oriented, it is possible to infer the cross-sectional profile of the tongue, which appears as a bright line in the resulting ultrasound image.

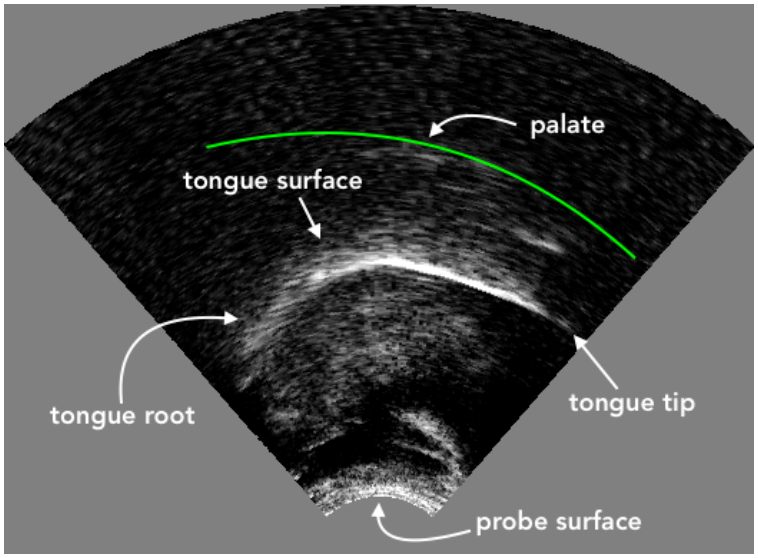

Figure 1: An ultrasound image of the tongue.

2.2 Electroglottography

Electroglottography (Fabre, 1957) is a technique that measures the size of contact between the vocal folds (the Vocal Folds Contact Area, VFCA). A high frequency low voltage electrical current is sent through two electrodes which are in contact with the surface of the neck, one on each side of the thyroid cartilage (Figure 2). The impedance of the current is directly correlated with VFCA, while its amplitude is inversely correlated (Titze, 1990). Thus, impedance increases with lower VFCA and decreases with higher VFCA. Conversely, amplitude decreases when the VFCA increases and it increases when the VFCA decreases. The EGG unit registers changes in impedance and it converts it in amplitude. Its output is a synchronised stereo recording which contains the EGG signal from the electrodes in right channel and the audio signal from the microphone in the left channel.

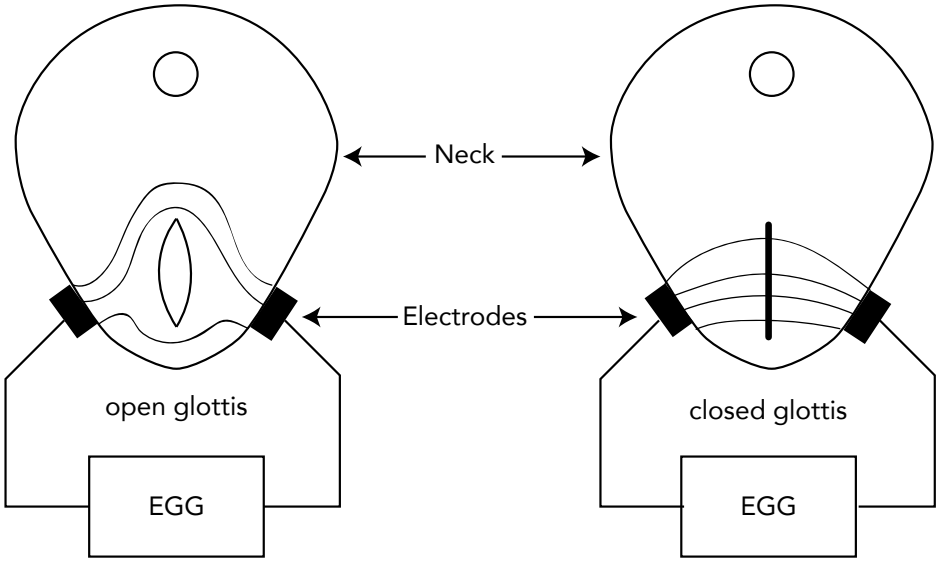

Figure 2: A schematic representation of the electroglottograph. The trasnversal section of the neck at the level of the glottis is represented: the effect of the glottis on the electric field are shown on the left for the open glottis and on the right figure for the closed glottis.

3 Methodology

This section describes the equipment setup, and the procedure for the acquisition, synchronisation, and analysis of the UTI and EGG data. Statistical analysis will be the subject of Section 4.

3.1 Equipment setup

Figure 3 shows the equipment set-up employed in the study. The left part of the figure shows the ultrasound set-up, while the EGG set-up is shown on the right. Two separate laptops are used for the acquisition of the ultrasound and EGG recordings. The ultrasound unit is plugged into one laptop. A P-Stretch unit (used for signal synchronisation) and the ultrasound probe are directly connected to the ultrasound unit. The P-Stretch unit and a microphone feed a pre-amplifier system, which is plugged into the ultrasound laptop. A second microphone and the electrodes are connected to the EGG unit, which is plugged into the second laptop. A TELEMED Echo Blaster 128 system is used for ultrasonography and a Glottal Enterprises EG2-PCX2 unit for EGG. The subject wears a headset (not shown in Figure 3) which holds the ultrasound probe in position (allowing free head movement) and a velcro strap with the EGG electrodes, located on each side of the thyroid cartilage, at the level of the glottis. The microphones are clipped to the headset on either side, at identical height.

3.2 Acquisition of ultrasound and EGG

Ultrasound and EGG inputs are acquired and recorded in separate laptops by means of, respectively, Articulate Assistant Advanced (AAA, Articulate Instruments Ltd, 2011) and Praat (Boersma & Weenink, 2016).

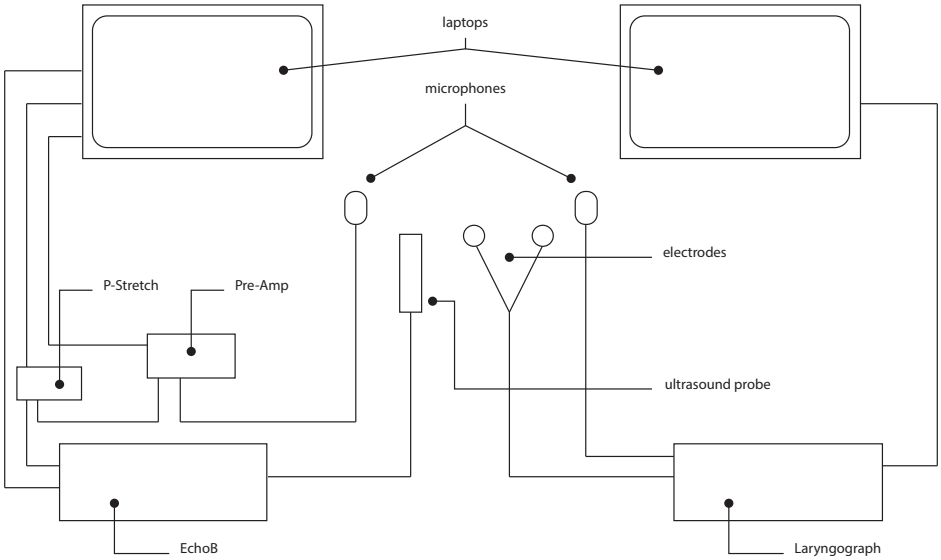

Figure 3: Equipment set-up scheme.

3.3 Ultrasound and EGG synchronisation

Since the signals from the ultrasound machine and the electroglottograph are recorded simultaneously but separately, data from both machines need to be synchronised after acquisition. Synchronisation is achieved through the cross-correlation of the audio signals from both sources (Grimaldi et al., 2008). This method creates a sound file from two audio files which is the cross-correlation of the original files (Figure 4). The interval between the start of the cross-correlated sound file and the time of maximum amplitude is equal to the lag between the two original files. Syncing of the original sound files is achieved by trimming the longer sound from its start by the same amount as the lag. A measure taken at any particular time in the ultrasound source can thus be related to a measure taken at that same time in the EGG source.

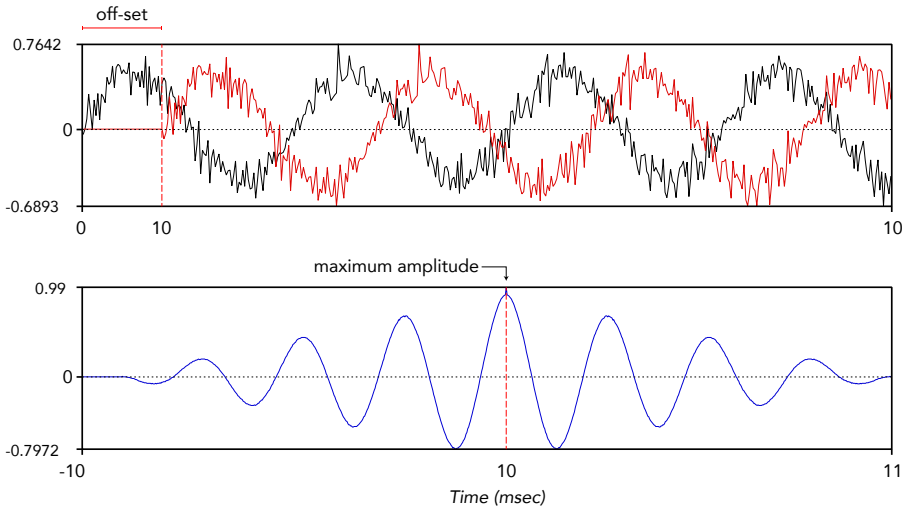

Figure 4: Synchronisation of two sounds by cross-correlation.

3.4 Analysis of ultrasound data

The ultrasound data is analysed with the same programme used for recording (Articulate Instruments Ltd, 2011). The tracking of the tongue contours is performed using a built-in automatic spline tracker. The splines as resulting from the automatic tracking are then checked and corrected manually if necessary. Velocity measures are subsequently calculated based on tongue displacement along two fan lines (one for the tongue tip and one for the tongue dorsum). Appropriate fan line selection is achieved by inspecting the standard deviation of tongue displacement on a few fan lines which lie on relevant portions of the tongue (front for the tongue tip and back-centre for tongue dorsum). The fan line of each respective tongue portion with the largest standard deviation is chosen and the velocity of tongue displacement along that line is calculated. Finally, four gestural landmarks (the onset and offset of the consonantal gesture and of its nucleus) are identified from the absolute velocity values using the method described in the following paragraph.

A single consonantal gesture is normally constituted by a closing phase that starts during the preceding vowel, a moment of maximum constriction during the consonant closure, and an opening phase that that terminates during the following vowel. The time of maximum constriction corresponds to the time of the minimum absolute velocity within the interval of the consonant closure (Figure 5). The velocity minimum is preceded and followed by two velocity maxima. The onset and offset of the nucleus of the consonantal gesture then correspond to the time where the absolute velocity reaches the 20% of the peak velocity, respectively before and after the maximum constriction. Finally, the gesture onset and offset are defined as the time where the peak velocity preceding and following the peaks reaches the 20% of the local peak velocity. The time and spline coordinates at each of the four gestural landmarks are then extracted for statistical analysis.

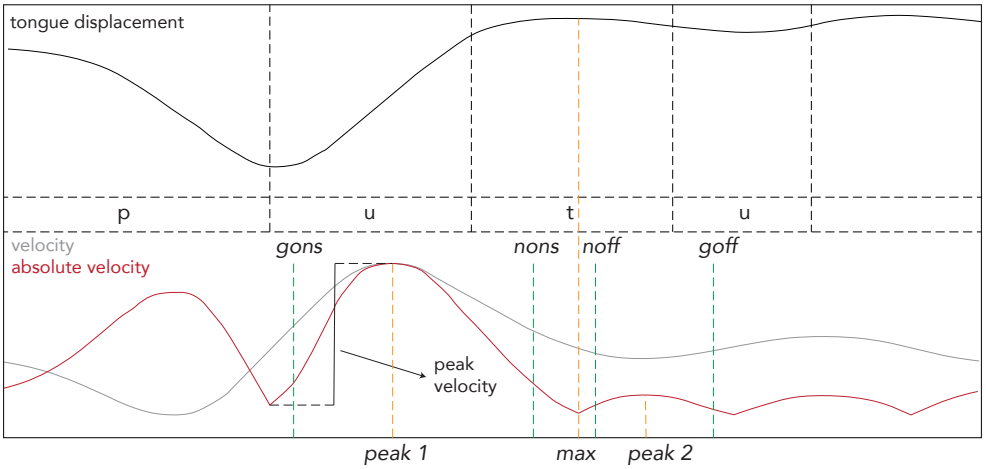

Figure 5: Identification of the gestural landmarks from the absolute tongue velocity profile (`gons` = gesture onset, `nons` = nucleus onset, `max` = maximum constriction, `noff` = nucleus offset, `goff` = gesture offset).

3.5 Analysis of EGG data

Previous studies have shown that the mathematical first derivative of the EGG signal helps determine the moments of glottal closure and opening in each vibration cycle, or glottal period (Henrich et al., 2004, 2005). The first derivative of a signal is the velocity of the signal, in other words how fast the signal changes in time. The time of maximum velocity in the first derivative of the EGG signal (dEGG) roughly corresponds to the

moment of glottal closure. The time of minimum velocity corresponds to the moment of glottal opening. Thus, glottal closure and opening for each glottal period can be extracted from the dEGG.

Herbst et al. (2010) describe a new technique, called electroglottographic wavegram, which displays the variations in the EGG and dEGG signals in a single graph. A wavegram contains temporal information on the *x* and *y* axis, while changes in the VFCA are rendered as different colour intensities on the *z* axis.

The extraction of dEGG maxima and minima has been implemented in this study using the PRAAT scripting language. The algorithm consists of the following stages:

1. detection of the glottal periods
2. calculation of the dEGG
3. extraction of absolute dEGG maximum ($dEGG_{max}$) and minimum ($dEGG_{min}$) for each glottal period
4. calculation of $dEGG_{max}$ and $dEGG_{min}$ relative to the glottal period

It is conventional to define a glottal period as the time between two consecutive moments of glottal closure, i.e. two consecutive dEGG maxima. However, since the maxima need to be identified in the first place, an arbitrary definition of glottal period is instead used. Glottal periods correspond to the intervals between two consecutive EGG minima [cf. ...]. First, the EGG signal is band-pass filtered (40Hz-10KHz) and smoothing is applied. A weighted sliding-average smoothing method (triangular smooth) is used, with smooth width *m* = 11. EGG minima are thus extracted from the smoothed EGG signal. The interval between any two consecutive minima constitutes a glottal period.

The dEGG is calculated with the formula $x'_n = x_{n+1} - x_n$, where x_n is the value of the EGG signal at the time n. After calculation, the resulting dEGG is smoothed with the same method as before (triangular smooth, *m* = 11). The algorithm then searches for dEGG maxima and minima within each glottal period (defined as two consecutive EGG minima). Finally, relative $dEGG_{max}$ and $dEGG_{min}$ are calculated as proportions of the respective glottal period. The resulting values are between 0 (beginning of period) and 1 (end of period).

As Herbst et al. (2010) note, the wavegram technique has the limitation of not being suitable for quantitative analysis. A new visualisation technique, based on wavegrams, is introduced here: electroglottographic tracegram. The tracegram method, even if it reduces the displayed dimensions, allows a statistical assessment of the varying $dEGG_{max}$ and $dEGG_{min}$, thus constituting a partial improvement over wavegrams. After the calculation of the relative $dEGG_{max}$ and $dEGG_{min}$, these values are plotted in a graph on the *y* axis at each time point which corresponds to the beginning of a glottal period. Since the values are restricted between 0 and 1 (being proportions), changes in glottal period (which corresponds to changes in fundamental frequency and hence pitch) are controlled for. The resulting graph, the tracegram, shows the traces of $dEGG_{max}$ and $dEGG_{min}$ as they change in time, in a way similar to the display of pitch contours.

4 Statistical analysis

Traditionally, differences in tongue contours has been assessed using Smoothing Spline ANOVA (SSANOVA) models (Gu, 2013; Davidson, 2006). More recently, generalised additive models (GAMs) have been applied to dynamic linguistic data including a space or time dimension (Wood, 2006; Sóskuthy, 2017). GAMs are considered to be a more conservative type of statistical model than SSANOVA, thus reducing the probability of getting false positive results. I am thus currently testing the application of GAMs on UTI and EGG data. Data post-processing and statistical analysis is performed in R (R Core Team, 2015).

5 Pilot study

I ran a pilot study to verify the performance of the methodology. Two speakers of Italian (2 males) and two speakers of Polish (1 female, 1 male) were recorded using the setup described in Section 3.1. The target words were of the form $C_1V_1C_2V_1$, where C_1 = /p/, V_1 = /a, o, u/, C_2 = /t, d, k, g/. The words were embedded in prosodically similar sentences (Italian *Dico X lentamente* 'I say X slowly', and Polish *Mówię X teraz* 'I say X now').

5.1 Results

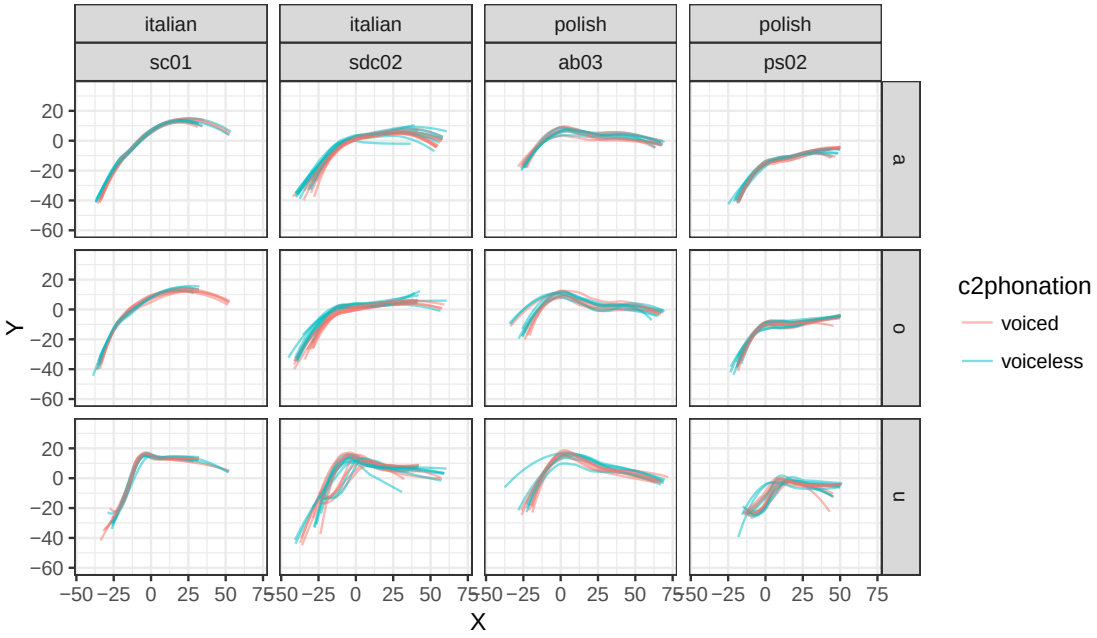

Figure 6: Tongue contour at maximum constriction of coronal consonants in Italian and Polish.

Figure 6 shows the tongue contours of the speakers of Italian and Polish. Only the contours at maximum closure of coronal consonants are plotted, separately for each vowel (rows) and speaker (columns). The root of the tongue is on the left of each graph, while the tip is on the right. The contours of the anterior part of the tongue in voiced (red lines) and voiceless stops (blue lines) overlap both in Italian and Polish. This indicates that the the sagittal shape of the tongue at maximum constriction is not affected by the voicing of the consonant. However, the tongue root seems to be advanced in voiced stops (red lines) compared to voiceless stops (blue lines) in Italian, but not in Polish. According to a generalised additive mixed effects model fitted on the Italian and Polish data, the tongue root was significantly advanced in voiced stops in Italian, while there was no significant difference in Polish (Figure 7).

The analysis of the EGG data is underway, although, as derived from the visual inspection of the data, it seems that both Italian and Polish voiceless stops entail earlier glottal spread, reflected in the tracegram of

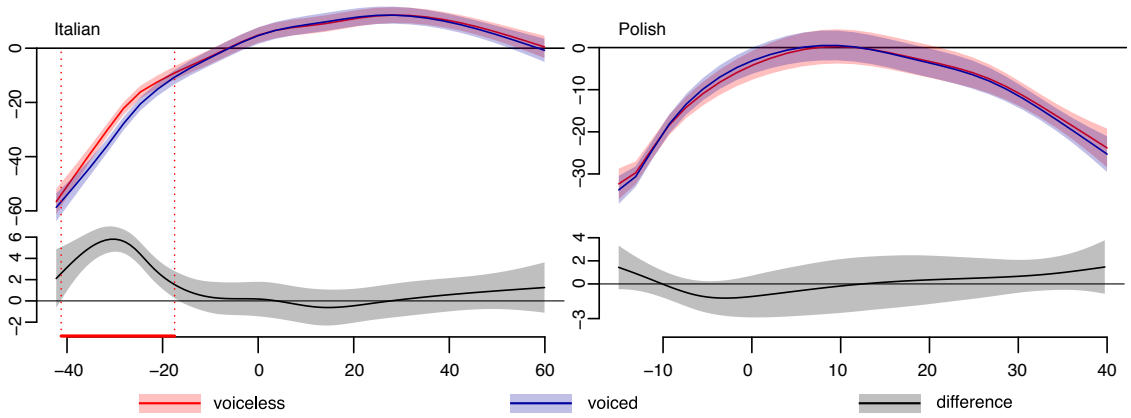

Figure 7: Tongue contour in voiceless and voiced stops at maximum constriction in Italian (left) and Polish (right). When the confidence intervals of the difference smooth (grey) do not cross 0 on the y-axis, they indicate a significant difference (red line on the x-axis).

these stops, shown in Figure 8. The dEGG maximum starts raising at around 60% of the vowel in Italian and Polish if the vowel is followed by voiceless stops. A raised dEGG maximum during the last portion of the vowel indicates that glottal spreading (which is the configuration of the vocal folds during the production of voiceless stops) is initiated even before consonant closure is achieved. On the contrary, there seems to be no significant increase in the dEGG maximum in voiced stops in neither language.

References

Articulate Instruments Ltd. 2011. Articulate Assistant Advanced user guide. Version 2.16.

Boersma, Paul & David Weenink. 2016. Praat: doing phonetics by computer [Computer program]. Version 6.0.23. http://www.praat.org/.

Davidson, Lisa. 2006. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America* 120(1). 407–415.

Fabre, P. 1957. Un procede electrique percutane d'inscrition de l'accolement glottique au cours de la phonation: glottographie de haute frequence. Premiers resultats. *Bulletin de l'Académie nationale de médecine* 141. 66.

Grimaldi, Mirko, B. Gili Fivela, Francesco Sigona, Michele Tavella, Paul Fitzpatrick, Laila Craighero, Luciano Fadiga, Giulio Sandini & Giorgio Metta. 2008. New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph. *Proceedings of LangTech* 1–5.

Gu, Chong. 2013. *Smoothing spline anova models*, vol. 297. Springer Science & Business Media.

Henrich, Nathalie, Christophe d'Alessandro, Boris Doval & Michele Castellengo. 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). 1321–1332.

Henrich, Nathalie, Christophe d'Alessandro, Boris Doval & Michele Castellengo. 2005. Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *The Journal of the Acoustical Society of America* 117(3). 1417–1430.

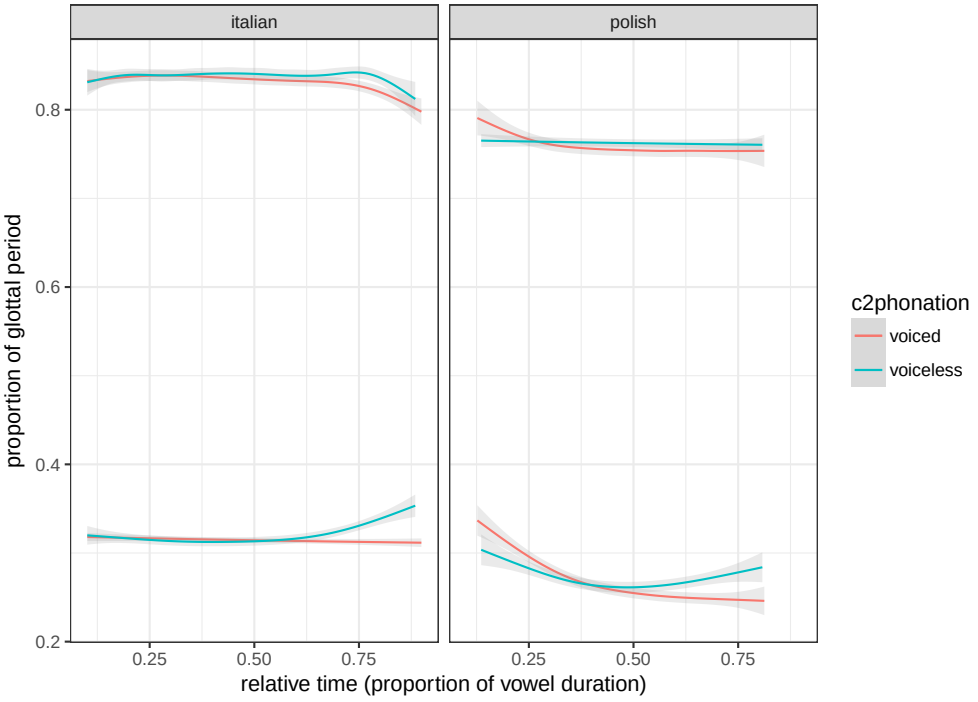

Figure 8: Tracegram.

Herbst, Christian T., W Tecumseh S. Fitch & Jan G. Švec. 2010. Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasively. *The Journal of the Acoustical Society of America* 128(5). 3070–3078.

R Core Team. 2015. R: A language and environment for statistical computing. https://www.R-project.org.

Sóskuthy, Márton. 2017. Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. arXiv preprint arXiv:1703.05339.

Titze, Ingo R. 1990. Interpretation of the electroglottographic signal. *Journal of Voice* 4(1). 1–9.

Wood, Simon. 2006. *Generalized additive models: an introduction with R*. CRC press.