Exploring Machine Learning perspectives for electroglottographic signals

Minh-Châu Nguyễn

CLD2025 – **C**omputational **L**anguage **D**ocumentation by **2025**

Work Package 2: Semi-automatic tonal models

Exploring Machine Learning perspectives for electroglottographic signals

Author:
Minh-Châu NGUYÊN

Project starting date: January 1st, 2022
Duration: 12 months
WP1 coordination: LIG
Contributors: CNRS-LIG, CNRS-LACITO

Revision: June 7, 2023

PRCI - International ANR-DFG		
Niveau de dissémination		
PU	Public	X
PP	Restreint aux autres participants du programme (y compris ANR et DFG)	
RE	Restreint à un groupe spécifié par le consortium (y compris ANR et DFG)	
CO	Confidential, restreint aux members du consortium (ANR et DFG)	

Abstract

LIG, CNRS

Université Grenoble Alpes

**Automatic Speech Recognition for less-studied languages:
Exploring Machine Learning perspectives for electroglottographic signals**

by Minh-Châu NGUYÊN

Studying and preserving under-resourced and endangered languages, for which few resources are available, is an arduous endeavour, which involves tremendously time-consuming tasks. In addition, for best results down the line (when the datasets are exploited for a variety of purposes), consistency in annotation work is highly desirable, but not easy to ensure in a manual workflow. High hopes are placed in innovative uses of speech recognition tools to facilitate and accelerate linguistic tasks.

Automatic Speech Recognition (ASR) is making remarkable progress thanks to the advent of deep neural networks (DNNs). A breakthrough is now possible: machine learning tools have improved to a point where they can effectively help to perform linguistic annotation tasks such as automatic transcription of audio recordings, automatic glossing of texts, and automatic word discovery. However, beyond that, there exists a considerable space for computer-assisted exploration and analysis of phonetic and phonological properties of languages. In this context, we attempt to explore the applicability of a neural network for phonetic/phonological analysis of an audio and electroglottographic corpus that has been previously processed manually. The objective of this work is to test the capabilities of neural networks to learn and reproduce specific strategies and principles for the estimation and manual verification of two phonetic parameters, namely fundamental frequency and glottal open quotient, which are acoustic correlates of pitch and phonation type (both of which constitute relevant dimensions of linguistic tone in the target language). In order to evaluate this learning process, a comparison of automatic and manual results is carried out.

This pilot study uses a recently collected and manually analyzed corpus[1] of the Kim Thuong dialect of Muong, a Vietic language that has a phonetically complex tonal in which one tone involves a lapse into creaky voice (M.-C. Nguyen, 2021). The acoustic and electroglottographic signals were recorded simultaneously. The manual annotation was based on the acoustic signal. The measurement of fundamental frequency and glottal open quotient was based on peak detection on the derivative of electroglottographic signal (also known as the DEGG signal) using the semi-automatic Peakdet script running on Matlab. Data from twenty speakers with twelve minimal sets plus three minimal pairs recorded within a carrier sentence amount to a total of five hours of recordings (an average of 18 minutes per speaker).

The results point to the presence of confounders, which (paradoxically) include the use of a carrier sentence: placing the target items in the same phonetic context stabilizes their tonal realizations, but the artificial neural network is biased towards easier predictions. But analysis of the results has benefits for being more explicit on the goals of analysis of electroglottographic signals, offering pointers for further work.

Keywords: language documentation, unwritten language, natural language processing, machine learning, neural networks, phonation types, creaky voice

[1] The corpus is available from the Pangloss Collection at https://pangloss.cnrs.fr/corpus/Mường under a Creative Commons license (CC BY-NC-SA 3.0 fr).

Acknowledgments

The work presented here is funded by the French-German project "Computational Language Documentation by 2025 / La documentation automatique des langues à l'horizon 2025" (CLD 2025, ANR-19-CE38-0015-04), conducted by an interdisciplinary team associating linguists and computer scientists (from the field of Natural Language Processing). Many thanks to Gilles Adda (project coordinator), Laurent Besacier and Alexis Michaud for initiating this project, which brought me this treasured opportunity to collaborate at LIG with the wonderful colleagues of the GETALP team to explore more possibilities to approach my data from a NLP perspective. It has been a great start along the long-term path of studying and preserving under-resourced languages with state-of-the-art computer tools. I greatly appreciated Alexis Michaud's participation and follow-up during the course of the present work and his careful proofreading of this report.

Many thanks to Séverine Guillaume and her team at LACITO for assistance with archiving the materials used in this study: materials from my Ph.D. research, archived in the Pangloss Collection, and used here for new research perspectives.

I am very grateful to Nathalie Henrich for encouraging Alexis Michaud and me to get in touch with Thomas Herbst. The exchange with him about new approaches to electroglottographic signals was extremely illuminating, suggesting several mid-term prospects that the present report hardly begins to address.

A big Thank You to Solange Rossato for her role as a bridge between the linguists and the computer scientists of the team thanks to her countless precious ideas, discussions and exchanges. Her advice has been very valuable and essential to the realization of this work. Many thanks also for her help in the statistical analysis with R for the data used in this project.

Last but not least, deep thanks to Maximin Coavoux, a core member of the team, who guided all the work on machine learning. Many thanks for his interest in this project, for taking the time and effort to tutor me starting at the very basics of natural language processing and deep learning, for always being supportive and available to answer my questions all along, up to the stage of proofreading this report. No words can express my deep gratitude to him.

Many thanks to all the members of the GETALP team for their warm welcome and support during the wonderful time I spent at LIG on this project.

Contents

1 Introduction **9**

1.1 Context: "Computational Language Documentation by 2025" . . . 9

1.2 Goals . . . 10

1.3 Electroglottography: principles, analysis methods, and prospects for automation of analysis processes . . . 10

1.3.1 Principles . . . 11

1.3.2 Methods to analyze the electroglottographic signal . . . 11

1.3.3 Criticism of estimation of the glottal open quotient by electroglottography . . . 13

2 Input: corpus **16**

2.1 The corpus: content (speech materials) and status of the data . . . 16

2.2 Note on some irregularities in the corpus . . . 17

2.3 Manual workflow . . . 19

2.3.1 General process . . . 19

2.3.2 Fundamental frequency measurement . . . 19

2.3.3 Glottal open quotient measurement . . . 20

3 $O_{q\ dEGG}$ estimation using machine learning methods **26**

3.1 Data Description and Preprocessing . . . 26

3.2 Neural Network Architecture . . . 27

3.3 Experiments . . . 28

3.4 Results and discussion . . . 29

Appendices **36**

A Detailed information on the speech material and the total corpus **37**

A.1 Speech material 37

A.2 Corpus status 42

Bibliography **44**

List of Figures

1.1 Example of EGG and dEGG signals 12

1.2 Continuum of phonation types 13

1.3 Procedure of creating an EGG wavegram 15

2.1 Calculation of the size of the total corpus. 17

2.2 A brief summary view of the corpus. 17

2.3 Basic procedure of data processing. 20

2.4 A schematic representation of data processing with PeakDet. 21

2.5 Examples of imprecise closing peaks 22

2.6 Examples of precise and imprecise opening peaks on DEGG signal 23

2.7 Example shows the role of manual $O_{q\ dEGG}$ verification: "good" multi-opening peaks . . 25

3.1 The neural network architecture 29

3.2 $O_{q\ dEGG}$ verification: case where methods of barycenter must be chosen 32

3.3 $O_{q\ dEGG}$ verification: case where methods of local minimum must be chosen 33

List of Tables

3.1 Statistics on the corpus 27
3.2 Final results on development and test sets (%). 30
3.3 Confusion matrix 30

A.1 Speech materials: minimal sets for tones in smooth syllables 38
A.2 Speech materials: minimal sets for tones in checked syllables 41
A.3 Summary of the processing status of the entire corpus 42

Abbreviations

dEGG First derivative of the electroglottographic signal

EGG Electroglottography, *or* electroglottographic, as in "EGG signal"

f_0 Fundamental frequency of speech

$f_{0\ dEGG}$ Fundamental frequency as estimated from the derivative of the electroglottographic signal

O_q Glottal open quotient

$O_{q\ dEGG}$ Glottal open quotient as estimated from the derivative of the electroglottographic signal

UID Unique Identifier

Chapter 1

Introduction

1.1 Context: "Computational Language Documentation by 2025"

The work reported here was carried out within the framework of the research project entitled "Computational Language Documentation by 2025" (hereafter, the CLD2025 project). The main objective of the CLD2025 project is to facilitate the urgent task of documenting endangered languages by leveraging the potential of computational methods. To recapitulate the core argument of the project application: until a decade ago, attempts at using Automatic Speech Recognition for low-resource languages (including newly documented languages) yielded modest results: there were interesting developments, but practical usefulness remained limited, and deployment as part of language workers' workflows still appeared as a prospect for the future (Besacier et al., 2014; Do, Michaud, and Castelli, 2014). A breakthrough is now possible: machine learning tools (such as artificial neural networks and Bayesian models) have improved to a point where they can effectively help to perform linguistic annotation tasks such as automatic transcription of audio recordings, automatic glossing of texts, and automatic word discovery (Thieberger, 2017; Michaud, Adams, Cohn, et al., 2018; Anastasopoulos et al., 2020). Significant achievements in this space include Partanen, Hämäläinen, and Klooster (2020), Prud'hommeaux et al. (2021), Liu, Spence, and Prud'hommeaux (2022), Macaire et al. (2022), and Rodríguez and Cox (2023).

The CLD 2025 project, "Computational Language Documentation by 2025", is organized in six work packages. The present task contributes to the second work package: "Semi-automatic tonal models", which aims at designing workflows involving automatic speech recognition tools to supersede purely manual workflows. In other words, the ultimate goal here is to contribute to the implementation of processing chains for the documentation of low-resource languages, which integrate automatic speech recognition tools.

Beyond the staple tasks of Natural Language Processing, there exists a considerable space for computer-assisted exploration and analysis of phonetic and phonological properties of languages. New strands of interdisciplinary research include linguistic reflections based on error analysis: reflecting on unexpected output from NLP tools, which offer a fresh perspective on the data (see e.g. Michaud, Adams, Cox, et al. 2020). Another strand consists in using machine-learning-assisted analysis of signals to obtain phonetic parameters to explore the phonetic and phonological characteristics of a language or dialect. The present work therefore aims to contribute to these new and challenging approaches.

1.2 Goals

A first-level goal of the present investigation is to automate the extraction of the glottal open quotient from electroglottographic recordings (some details on this technique will be provided in section 1.3). That is a time-consuming annotation task which requires a certain degree of expertise. Based on my experience of linguistic fieldwork, I concur with the observation that computer-assisted work is highly desirable over fully manual workflows. In this respect, the annotation of electroglottographic signals is similar to standard tasks in Natural Language Processing, such as speech recognition, translation, and glossing. Manual annotation and semi-automatic analysis of electroglottographic signals using the PeakDet tool (as described in M.-C. Nguyen 2021, pp. 99–110) took me roughly 15 months to carry out for the data of 20 speakers. The recording duration for each speaker's data is just about 15 minutes, and the corpus has a simple structure: the 66 target syllables were embedded inside a carrier sentence (details are provided in Chapter 2 below). Performing the same work on a corpus of spontaneous speech would be even more complex and tedious in view of the well-documented variability of spontaneous speech. Moreover, workflows based on manual decisions are not technically reproducible, raising thorny epistemological issues. It would clearly not be desirable for electroglottographic analysis to become a craft proudly performed by a guild of experts on the rather vague basis of personal experience. Such a situation would be reminiscent of the problem with Cardinal Vowels, for which the ultimate reference was personal instruction from Daniel Jones... and notable variation is found from one generation of students to the next, defeating the purpose of this set of reference vowels (Vaissière, 2011).

Conversely, under the hypothesis (which, as we shall see below, was clearly over-optimistic) that a model could be trained to replicate the task of analysis of the electroglottographic signal, then the entire set of available (open-access) electroglottographic recordings on a range of typologically diverse languages could be processed in a consistent and reproducible manner, opening new avenues for the cross-linguistic study of phonation types in human speech, in full-reproducibility mode – a highly desirable development for the field of speech sciences (Kobrock et al., 2023).

My contribution consisted in: (i) preparing and managing linguistic data for statistical processing and for a pilot study on the application of a neural network for the estimation of the glottal open quotient; (ii) writing this technical report as a reference for linguists (non-computer scientists) in approaching and applying machine learning tools to data processing and archiving.

The work was carried out at LIG (Laboratoire d'Informatique de Grenoble, UMR 5217), under the joint supervision of a linguist, Solange Rossato, and a computer scientist, Maximin Coavoux (in collaboration with Alexis Michaud, from LACITO, UMR 7107).

1.3 Electroglottography: principles, analysis methods, and prospects for automation of analysis processes

In this section, we present a short introduction to the collection and analysis of electroglottographic data. The technical explanations here aim to summarize the underlying principles of electroglottography, and the ways in which electroglottographic signals can be used in research. This part is crucially important because the linguist must communicate well with the computer scientist who will take over the data to train them in the neural network. The more precise and elaborate explanation is provided about the data and its usefulness to research, the better a computer scientist can understand the targets of the task and come up with solutions for automation. As an example: among the limitations acknowledged in Section 3.4, we will discuss an instance of misleading communication on my part that affected the

outcome of the study.

1.3.1 Principles

The electroglottographic signal provides an estimate of variation in the contact area between the two vocal folds. Electroglottography (often abbreviated to EGG) was invented by Fabre in the mid-20th century. The initial report about the invention Fabre (1957) was followed by further studies by the same author over the following years (Fabre, 1958; Fabre, 1959; Fabre, 1961), initiating strands of research that are still active to this day.

> Electroglottography is a common, widespread technique that enables the investigation of vocal-fold contact area in phonation in an easy and noninvasive way. A high frequency modulated current (F = 1 MHz) is sent through the neck of the subject. Between the electrodes, electrical admittance varies with the vibratory movements of the vocal folds, increasing as the vocal folds increase in contact. (Henrich et al., 2004, p. 1321)

The EGG signal is a continuous signal, like the audio signal. It can therefore be stored in the same format as the audio, and displayed with the same tools.

1.3.2 Methods to analyze the electroglottographic signal

In the previous manual/semi-automatic processing, the method chosen for analysis of the electroglottographic signal uses its derivative signal (dEGG). Glottis-closure instants are approximated through detection of positive peaks in the first derivative of the signal, and glottis-opening instants through detection of negative peaks in-between the positive peaks.

This method is set out in full by Henrich et al. (2004). Henrich's paper goes into technical detail concerning the four main phases of a glottal cycle: (i) closing phase, (ii) closed phase, (iii) opening phase, and (iv) open phase. Increase in vocal fold contact area is reflected by the closing phase (itself followed by the closed phase) in the electroglottographic signal, and the moment of fastest increase in vocal fold contact area corresponds to the glottis-closure instant. Decrease in vocal fold contact area begins during the closed phase, and continues into the opening phase; the moment of fastest decrease in vocal fold contact area is the glottis-opening instant.

But the correspondence between these four main phases, on the one hand, and detectable events on the electroglottographic signal, on the other hand, is not easy to establish. Instead, the electroglottographic signal corresponding to one glottal cycle can be divided into two portions only, as shown on Figure 1.1. These two portions are named *closed phase* and *open phase* of the vocal-fold vibratory cycle, and defined as follows: the closed phase extends from a glottis-closure instant to the next glottis-opening instant; and the rest of the cycle is the open phase (for a more detailed description, see Henrich et al. 2004, pp. 1321–1322, as well as D. G. Childers and Krishnamurthy 1984, Colton and Conture 1990, Orlikoff 1998.)

The derivative of the electroglottographic signal (dEGG) signal typically has a positive peak at glottis closure and a negative peak at glottis opening. Figure 1.1 illustrates visually a synchronization of EGG and dEGG signals. In the case of clear signals, one closing peak is clearly visible for each cycle, corresponding to the peak increase in vocal fold contact area and considered as the beginning of the glottal closed phase, and one (less salient) opening peak, corresponding to a peak in the decrease in vocal

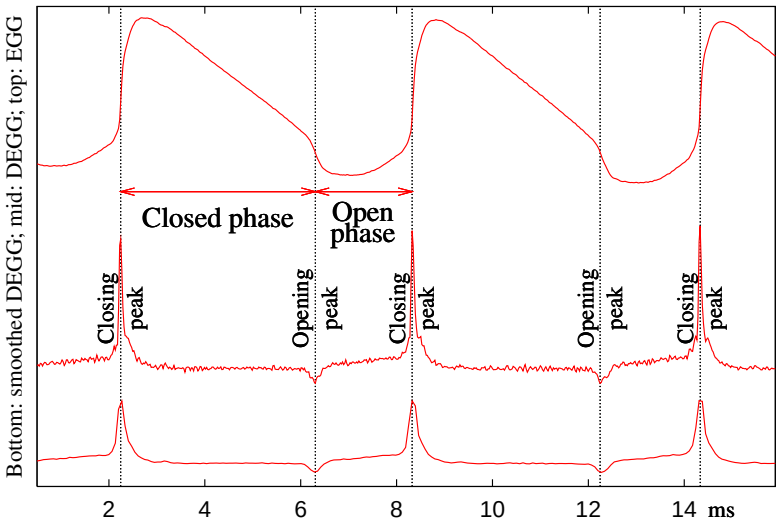


Figure 1.1: Example of EGG and dEGG signals with indication of glottis closure and opening. Reproduced with permission from the author, Alexis Michaud.

fold contact area and considered as the beginning of the glottal open phase. These peaks in the dEGG signal serve as the basis for estimating glottal parameters. The most well-known parameter is speech fundamental frequency (f_0). The glottal open quotient is less well-known among linguists, as it cannot be easily estimated from the audio signal only, but it is used in many linguistic studies of phonation types and tones: see in particular Michaud (2004b), Brunelle, Nguyên, and K. H. Nguyen (2010), Abramson, Tiede, and Luangthongkum (2015), Brunelle, Tấn, et al. (2020), and Kirby, Pittayaporn, and Brunelle (2022). Additionally, the amplitude of the closing peak (Derivative-Electroglottographic Closure Peak Amplitude, abbreviated as DECPA) can be measured from the dEGG signal (Michaud, 2004a; Kuang and Keating, 2014); it is not presented and not studied here because the relationship of this parameter to phonation types is still not well established.

Fundamental frequency (f_0, unit: Hz) is the inverse of the glottal period (i.e. the inverse of glottal cycle duration). Specifically, $f_{0\ dEGG}$ is obtained by measuring the duration between two consecutive glottal closing instants, corresponding to a period (one glottal cycle). The inverse of the duration of the cycle yields the fundamental frequency of the voice (the formula is simple: F = 1 / T). The values of f_0 have *pitch* as their perceptual counterpart: low f_0 is heard as low pitch, and high f_0 as high pitch.

Glottal open quotient (O_q, unit: %). The measurement of $O_{q\ dEGG}$ requires the measurement of the duration of the glottal cycle, plus the detection of the glottal opening instant. This allows for computing the glottis-open interval; the open quotient is the ratio of the open-glottis interval to the entire cycle (the ratio between open time and fundamental period). This can be stated as the following equation: O_q = (Open phase) / (Open phase+Closed phase). O_q is a parameter that relates to phonation types: low O_q demonstrates pressed phonation; medium O_q reflects modal phonation; and high O_q reflects flow phonation (whispery voice, shading into breathy voice). This relates to the following observation:

> There might be a continuum of phonation types, defined in terms of the aperture between the arytenoid cartilages, ranging from voiceless (furthest apart), through breathy voiced, to regular, modal voicing, and then on through creaky voice to glottal closure (closest together). This continuum is depicted schematically in Figure 1.2. (Gordon and Ladefoged, 2001, p. 384)

A strong motivation for adopting the dEGG method for estimating O_q is comparability across studies. This method is fairly widely used in phonetic studies of phonation types published since Nathalie Henrich's methodological article (Michaud, Vu-Ngoc, et al. 2006; Mazaudon and Michaud 2008; Gao 2016, among others), and also in various other phonetic studies (e.g. Recasens and Mira 2013). The use of similar algorithms facilitates comparison across studies, and hence across languages as well as across speakers and across datasets.

Moreover, the dEGG method is grounded in explicit assumptions that relate to physiological observations in a way which, although not simple and straightforward, is intuitively clear.

1.3.3 Criticism of estimation of the glottal open quotient by electroglottography

Estimation of the glottal open quotient by electroglottography has come under criticism, which it appears useful to review here.

In a review article entitled "Electroglottography – an update", Herbst (2020) recapitulates important caveats about the interpretation of the electroglottographic signal. Some of them are well-known: "Vocal fold vibration, a complex phenomenon taking place in three spatial dimensions, is mapped onto a single time-varying value" (Herbst, 2020, p. 4). It needs to be borne in mind that electroglottography provides a linear insight into phenomena that are not linear, and thus only offers glimpses into complex phenomena, which ideally need to be addressed through an array of exploratory techniques: a multisensor platform (Vaissière et al., 2010).

But Herbst's criticism cuts deeper. He questions the assumption that underpins the method employed here: that peaks on the derivative of the EGG signal provide reliable estimates of the timing of glottis-closure instants. Reviewing recent studies, he considers that they "strongly suggest that positive and negative dEGG peaks do not necessarily precisely coincide with GCI (i.e. glottis closing instant) and GOI (i.e. glottis opening instant), a notion that was already put forward by D. Childers and Lee (1991), who maintained that the EGG signal *may not provide an exact indication for the instant of glottal closure*" (Herbst, 2020, p. 7). As emphasized by Hampala et al. (2016), "any quantitative and statistical data derived from EGG should be interpreted cautiously, allowing for potential deviations from true VFCA [Vocal Fold Contact Area]". The criticism is then extended to the very notions of glottis-closure instant and glottis-opening instant: "vocal fold contacting and de-contacting (as measured by EGG) actually do not occur at infinitesimally small **instants** of time, but extend over a certain **interval**, particularly under the influence of anterior-posterior (...) and inferior-superior phase differences of vocal fold vibration" (Herbst 2020, p. 7; emphasis in original).

From a theoretical point of view, there may be a slight confusion here, as surely no one among users of

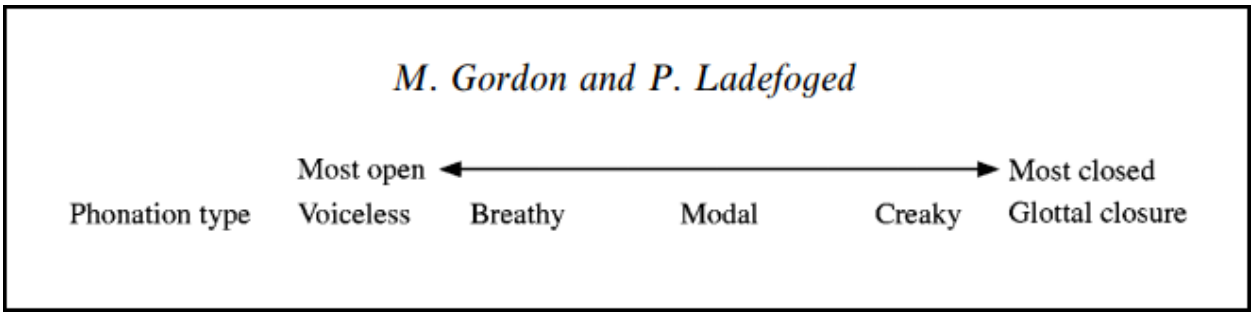


Figure 1.2: Continuum of phonation types (Reproduced from Gordon and Ladefoged, 2001, p. 384).

the method of estimating f_0 and O_q by means of peaks on the dEGG signal believes that glottal activity consists of instantaneous events of glottis closing and opening. The notions of glottis-closure instant and glottis-opening instant should, as a matter of course, be delivered complete with due precautions and careful hedging for their proper interpretation, but these precautions do not detract from the usefulness of these concepts. It should suffice to say once and for all that f_0 and O_q as estimated through the dEGG method should not be confused with the physical parameter that they aim to capture. A good way to make this distinction consistently consists in embedding a reminder about the estimation method within the acronym used for the parameter. Therefore, the notations $f_{0\ dEGG}$ and $O_{q\ dEGG}$ are adopted throughout the present report to refer to the measured parameters, as distinct from f_0 and O_q, the latter being understood either as abstract and ideal, or as generic labels.

From a practical point of view, a key point here is what is meant by "precisely" when claiming that dEGG peaks do not coincide "precisely" with glottis-closure instants and glottis-opening instants. The weak claim that "perhaps the glottal area waveform, if available, would be a more suitable candidate" than the dEGG signal as a ground truth for glottal events is perfectly safe as a hypothesis, but hardly helpful for those to whom the glottal area waveform is simply not available. In practice, the difficulty of obtaining low-noise electroglottographic signals is a much more serious subject of concern to me than the fully accepted theoretical limitation whereby "the determination of contacting and de-contacting instants or events is an artificial concept" (Herbst, 2020, p. 10). The fact that glottis-opening instants as estimated from dEGG signals may be slightly earlier than those obtained by other methods does not detract from cross-token, cross-speaker and cross-language comparability, and common sense suggests that those are precious assets.

On topics of terminology, Herbst's proposals are not particularly straightforward to implement. He uses 'closed quotient' (C_q) rather than 'open quotient' (O_q), which is not a real difference at all:

$C_q = 1 - O_q$

He argues that 'closed quotient' should be replaced by 'contact quotient':

> Given that the underlying EGG signal measures relative vocal fold **contact** area and not glottal closure, the terminology for that parameter should be limited to "**contact quotient**" instead of "**closed quotient**". Consequently, the term "open quotient" is also inappropriate, because EGG does not measure glottal opening. Instead, the term "quasi open quotient" (QOQ) might be used. (Herbst, 2020, p. 11)

I leave it to more established researchers to decide whether to take the turn towards use of the term "quasi open quotient" (QOQ). Trying to weigh the advantages, I find them very slight, compared to O_q. The suggestion to prefix "quasi" to the term "open quotient" strikes me as standing in contradiction to the statement (made by the author earlier on in the same paragraph) that this parameter "is not an *ersatz* closed quotient" (Herbst, 2020, p. 11). Among prefixes, "quasi" sounds like a reasonable equivalent for description as "*ersatz*": an inferior substitute or imitation, used to replace something that is unavailable and can only be approached, not equated.

Within studies related to electroglottography, "quasi" also brings to mind a proposal to build a "quasi-glottogram signal" from the EGG signal (Kochanski and Shih, 2003). While I can make no claim to understanding the maths sustaining the attempt to build a "quasi-glottogram signal", it is intuitively clear that the relationship between glottogram and electroglottogram in this proposal (published in the *Journal of the Acoustical Society of America*) was a much less straightforward one than that which links QOQ with O_q. To conclude, it does not seem completely fair or productive to dismiss O_q along with all other estimations of the glottal open quotient through electroglottography.

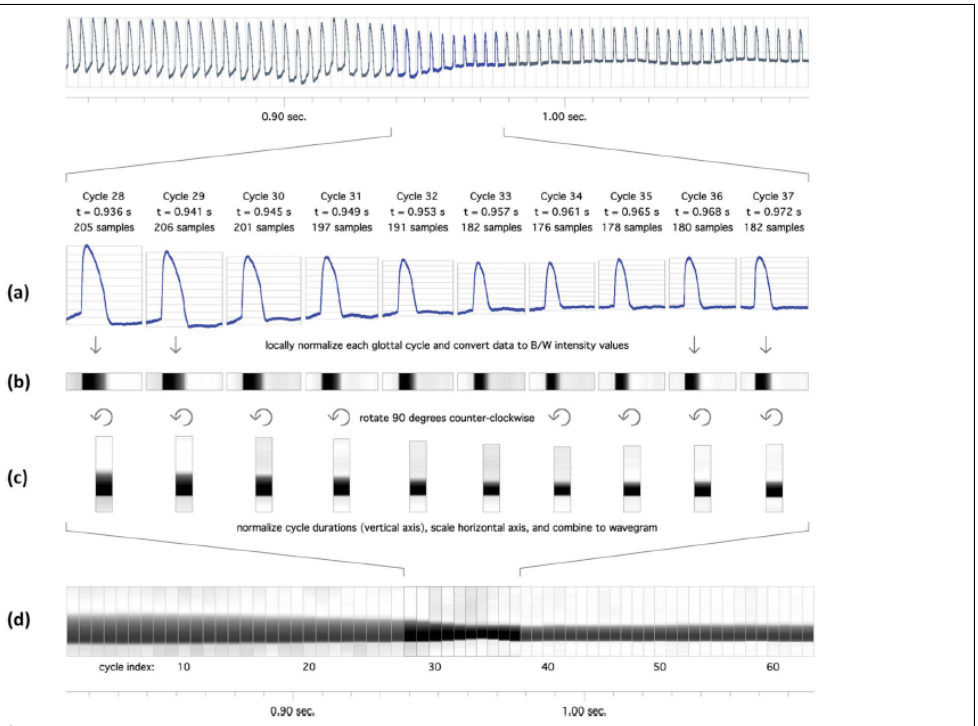


FIG. 3. (Color online) Illustration of the basic processing steps to create an electroglottographic wavegram: consecutive glottal cycles of an EGG signal are extracted (a). The locally normalized data values are converted into monochrome color information, and are plotted as a strip representing one glottal cycle each (b). All strips are rotated 90° counter-clockwise (c). Glottal cycle duration is normalized by scaling the individual glottal cycle plots to the same height, and the resulting graphs are combined to form the final display, the EGG wavegram (d), and the resulting graphs are combined to form the final display, the EGG wavegram (e).

Figure 1.3: Basic procedure of creating an electroglottographic wavegram. Reproduction from Herbst, Fitch, and Švec (2010, p. 3072)

Another technique proposed by Herbst, Fitch, and Švec (2010) for analyzing and displaying EGG and dEGG signals is named "wavegram". To construct a wavegram, the consecutive individual glottal cycles in EGG or dEGG signals are identified and extracted. They are locally normalized in duration and amplitude, then encoded by color intensity (monochrome color information). And finally, the color-coded strips, corresponding to glottal cycles, are rotated by 90° counter-clockwise and concatenated to display the entire voice sample in a single image. Figure 1.3 is reproduced from Herbst, Fitch, and Švec (2010, p. 3072) for an easier and more visual understanding of this method. According to the authors, the wavegram technique could provide a potentially powerful method for displaying entire electroglottographic signals, or parts thereof. Much like in the spectrogram, information on vibratory behavior developing in time is compacted into one single graph providing insight into changes of vocal fold dynamics. However, this method is more reliable for quasi-periodic phonation where the detection of glottal cycles can be easily determined by performing an auto-correlation analysis. The case of non-periodic phonation, on the other hand, is more complicated to address (to the point that it is doubtful whether this method can be made to succeed). In the article, this problem was also pointed out and they proposed an alternative algorithm for glottal cycle detection: "the period should be rather determined on a cycle-to-cycle basis from direct inspection of the electroglottographic signal and its first derivative in the time domain", i.e. falling back on a manual workflow.

The material of the current study involves creaky voice, a case of non-periodic phonation, as the target phenomenon. It could be interesting in the future to apply and test the wavegram in our data and compare it to the available results to see to what extent this technique can be applied to non-modal phonation and if it is a better approach to the EGG signal than the dEGG signal. For now, we retain the open quotient as the main glottal parameter extracted specifically from the electroglottographic signal.

Chapter 2

Input: corpus

2.1 The corpus: content (speech materials) and status of the data

The speech materials for this experiment is composed of minimal sets of real words. In total, it consists of 12 minimal sets[1] that contrast for tone in smooth syllables (i.e., open syllables or syllables ending with a nasal coda) plus 3 minimal pairs that contrast for tone in checked syllables (i.e., syllables ending with a stop coda). Tables A.1 and A.2 in Appendix A provide full detail about these minimal sets and pairs.

Method of collection: each target word is required to be spoken four times: twice in isolation and twice in a carrier sentence.

The carrier sentence is a question including 4 words:

(1) **/ja² măt⁶ ____ ʈăŋ³/**
2SG to_know *target item* INTERROG
'Do you know **_____**?'

The total corpus (per speaker): Figure 2.1 recapitulates the total corpus of this study. Not only the target words but also the three frame words of the carrier sentence are annotated and processed. Thus, for each speaker, we have a total of 660 items, of which 264 items are target syllables and 396 items are frame syllables. A more detailed list of the amount of materials is given in Table 2.2. In some cases, the maximum number of items is not reached because some frame words are missing, as speakers tend to shorten the carrier sentence during a series of repetitions. The most serious case is in the data of the speaker F10. For some technical reason, we made a pause but mistakenly did not press the record button to resume, so the last part of the experiment was missed on the first run. In particular, the minimal set N°11 in carrier sentence, the minimal set N°12 and all three minimal pairs both in isolation and in carrier sentence were not recorded. As a consequence, this data lacks 75 items, including 11 target words in isolation and 16 target words in carrier sentence, which leads to the sorely felt absence of 48 frame words at all three positions (i.e. 16 items for each).

[1] In 12 minimal sets, there are 8 complete minimal sets and 4 near minimal sets.

> "Pairs that show segments in nearly identical environments, such as azure/assure or author/either, are called near-minimal pairs. They help to establish contrasts where no minimal pairs can be found." (Dobrovolsky and Katamba, 1996).

((5 x 12) + (2 x 3) + (5 x 12 x 4) + (2 x 3 x 4)) x 2 = 660 (items)

Elements	Meaning
5	5 tones in smooth syllables
12	12 minimal sets (of 5 smooth tones)
2	2 tones in checked syllables
3	3 minimal pairs (of 2 checked tones)
4	4 syllables of the carrier sentence (1 target word + 3 frame words)
2	2 repetitions

Figure 2.1: Calculation of the size of the total corpus.

Total corpus: 660 items		
Target syllables: 264 items		**Frame syllables:** 396 items
In isolation: 132 items - 24 tokens each smooth tone (x 5 tones) - 6 tokens each checked tone (x 2 tones)	**In carrier sentence:** 132 items - 24 tokens each smooth tone (x 5 tones) - 6 tokens each checked tone (x 2 tones)	- /ɟa²/ : 132 tokens - /măt⁶/ : 132 tokens - /căŋ³/ : 132 tokens

Figure 2.2: A brief summary view of the corpus.

The actual status of the data of each speaker is summarized in Table A.3. There are a total of 26 participants, 28 data files (F1 and M12 performed the experiment 2 times), twenty of which have been processed.

2.2 Note on some irregularities in the corpus

Some asymmetrical points of the data There were some asymmetry defects from the data related to factors not accounted for in data collection, annotation, and pre-processing.

The most notable asymmetry in the data is the large difference in number between the sets for smooth tones and the pairs for checked tones. We have 12 minimal sets of smooth tones but only 3 minimal pairs of checked tones. Therefore, the target samples of the checked tones are four times fewer than those of the smooth tones. This is due to the fact that when designing the data collection, I did not consider the balance of the data between tones. I only considered syllables that were already found in the system of smooth tones, and then combined them with a final stop. This is unnecessary and limits the checked pairs that could be found. There are a few minimal pairs that were found but were omitted during the recording process due to inaccuracies in meaning or loanwords from Vietnamese. A detailed report on this topic is found in my Ph.D. dissertation at the end of the table of minimal pairs (M.-C. Nguyen, 2021, p. 67).

The second asymmetry is that the total number of items processed is not identical in the data of twenty speakers. As previously mentioned in Section 2.1, the data for each speaker normally contains a total of 660 items. However, due to a technical problems during the recording session and the tendency to shorten the carrier sentence during a series of repetitions in some cases, there are therefore a few words

missing from the data of 3 among the 20 speakers. The actual status of the data of each speaker is summarized in Table A.3. There are a total of 26 participants, 28 data files (F1 and M12 perform the experiment 2 times) of which 20 have been processed. Among the 20 data files that have been processed, 17/20 files have full size of 660 items, 2/20 files (the data of F13 and M14) have missing items from frame words, and 1/20 (the data of F10) file have missing items from both frame words and target words.

One of the data chunks that has the full size of 660 items but that nonetheless calls for special attention is that contributed by speaker F20. An issue with these data is that the ratio of excluded O_q values stands at 83%, which is especially high and makes it an outlier. The average of this ratio for the other speakers is 18%, with the highest ratio being 36.3% (for F9) and the lowest ratio being 4.7% (for M9). The reason for suppression of O_q values in speakers is often due to unclear opening peaks when the voice breaks into creaky voice. But this is not the case for F20. Not only the syllables carrying Tone 4 but all other tones have the same situation with imprecise opening peaks, making it impossible to measure O_q values. The consistent behavior of these peaks throughout the experiment until the last minimal set over the syllable **/ku/** (which was performed twice separately) provides evidence which makes me believe that this is related to a physiological phenomenon, rather than artifacts. This interesting case would be worth studying further and should be kept an eye on for this present study when we use the O_q values from the semi-manual process for machine learning process and also to evaluate the result later.

During the pre-processing to prepare data for this experiment and run it using machine learning methods, two unusual points have been detected from the previous manual processing.

Error on data of F3: The first unusual point was noticed during the process of preparing data by extracting MFCC frames. In the result of the speaker F3 from manual processing, there were some cases (11 cases, exactly) where the last annotated cycles of a syllable were outside its time interval.

In order to understand precisely this problem, we must first get to know the process explained later in the section 3.1. In short, for running the machine learning models, the required data are the EGG signal and the results from the manual processing which are stored in a three-dimension matrix in Matlab files. The necessary information extracted from Matlab matrices are stored in Excel files, including: the identifier and time intervals of each item (syllable), the starting and ending time of each glottal cycle detected within the syllable, and the corresponding values of $f_{0\ dEGG}$ and $O_{q\ dEGG}$ if measurable.

Logically, the sum of the duration of all detected glottal cycles inside a syllabe would be exactly the duration of that syllable. Therefore, the error was detected when, in some items, the last periods are out of range of the syllable duration. Thanks to automatic processing, such error can be easily detected.

This occurred only in the data from speaker F3 and there were 11 items, which are not consecutive, that had this error. Considering all these erroneous items, I found a pattern in which the spurious length of the erroneous items is duplicated from the item that follows immediately after them. For example, item with UID 0431 is one of the erroneous items. The incorrect length in the result is 41 glottal cycles, which is the correct lenght of the next item, which has UID 0432. By checking the autosave of each individual item, I figured out that the real length of item 0431 is 31 cycles instead of 41 cycles. The length is copied from item 3042, but the values of the extra 10 cycles at the end of the item 0431 are not the same as the ones at item 0432. This part is patched from somewhere else and actually does not correspond to any part in the data.

The solution for this error is that I can simply cut off the extra spurious cycles of 11 erroneous items. It was easy to check the erroneous part since the starting time of this part is not consecutive to the correct part.

I could not explain how this happened during the semi-manual processing. But by referring to the manual processing log, I can guess that this error occurred due to the fact that the data from F3 being annotated and processed twice because of some mistakes made during the first processing. And in the second time of processing, I did not start from the beginning, but from the token where I made the mistake for the first time. The part from the beginning to this token was resumed from the previous processing.

This shows a side benefit of computational tools, which is that it can help detect manual processing errors, which are difficult to check manually in a systematic way.

2.3 Manual workflow

In order to be better prepared and to have a better understanding of what the neural network has to learn automatically, this part tries to briefly recap the procedure of manual processing that produced the data which will be used down the line as a basis for tests with machine-learning tools.

In order to study the tone system of the target language, two phonetic parameters, fundamental frequency ($f_{0\ dEGG}$) and glottal open quotient ($O_{q\ dEGG}$), were estimated from the *derivative* of the EGG signal, DEGG (Henrich et al., 2004), using PeakDet, a script available from the COVAREP repository Degottex et al., 2014. (An implementation in Praat is also available: Kirby, 2020.) PeakDet is designed for semi-automatic measurement: the results for each token are verified visually.

2.3.1 General process

Figure 2.3 shows the basic procedure that was applied to the data set of twenty speakers to obtain results.

For the most part, the initial input materials for this process are audio files obtained from the minimal set experiment. They are first segmented and annotated (using the Sound Forge software) to obtain the annotated EGG file in mono-channel format and the Regions List indicating the time codes for each token, together with its unique identifier (UID). These are the two inputs required by PeakDet, a semi-automatic tool for estimating $f_{0\ dEGG}$ and $O_{q\ dEGG}$ from the EGG signals. This meticulous (and time-consuming) verification process is summarized in the algorithm shown as Figure 2.4.

2.3.2 Fundamental frequency measurement

In this study, the f_0 is estimated from the derivative of the electroglottographic signal in the EGG signal, hence we call it $f_{0\ dEGG}$. To measure f_0, the period is determined by detecting two successive closing peaks, the duration between two consecutive glottal closing instants corresponds to a fundamental period; its inverse gives the fundamental frequency of the voice. In the majority of cases, the closing peak is unique and precise in each glottal period to obtain the $f_{0\ dEGG}$. Figure 1.1 shows the ideal example of well-defined closing peaks. However, in a few cases, imprecise multi-closing peaks linked to a physiological phenomenon could be encountered. Figure 2.5 shows examples of double-closing peaks and jagged-shaped multi-closing peaks. In these cases, PeakDet is automatically set to use the barycentre method (see Section 2.3.3).

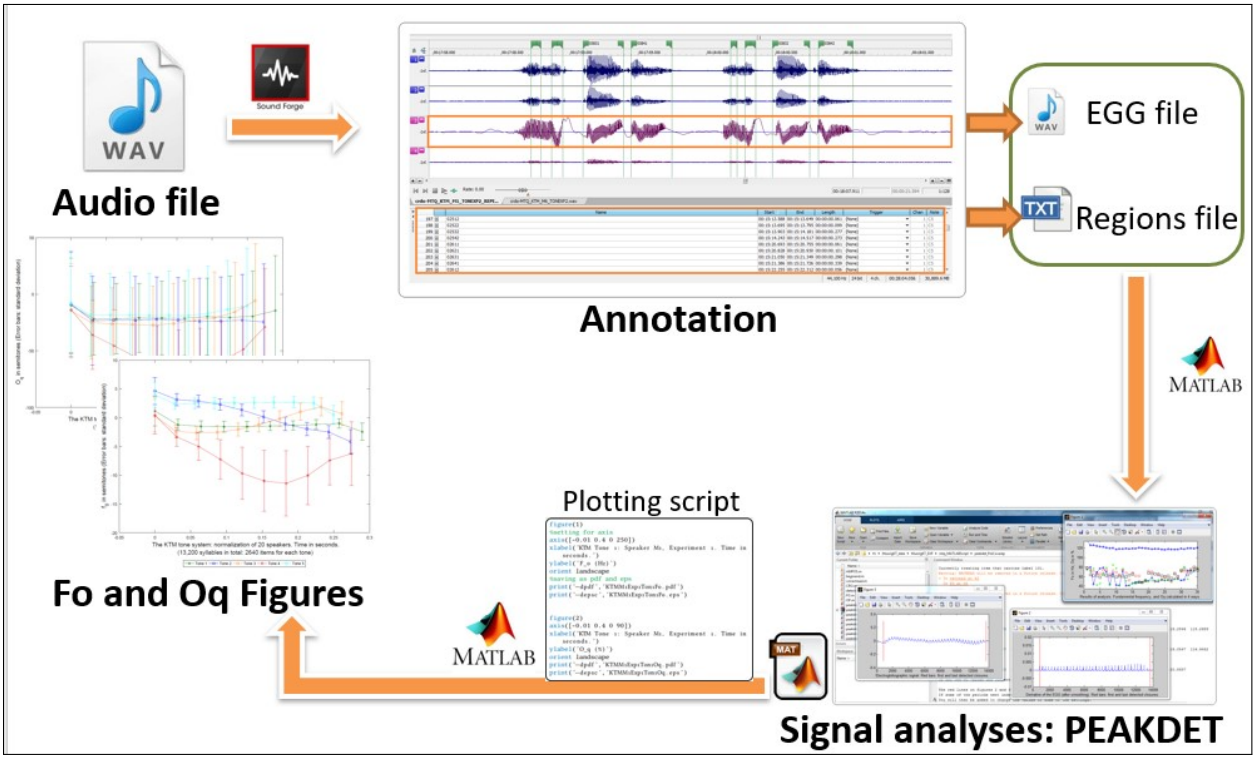


Figure 2.3: Basic procedure of data processing.

2.3.3 Glottal open quotient measurement

Whereas $f_{0\ dEGG}$ is calculated based on closing peaks, which in the great majority of cases are well-defined (often with a unique peak), estimating $O_{q\ dEGG}$ requires the detection of opening peaks, which often runs into difficulties due to imprecise peaks: either cases where no peak stands out clearly, or cases where two or more peaks are present (multiple peaks). The search for opening peaks is even more difficult in the case of nonmodal phonation, such as when voicing transitions into creaky voice.

This makes user verification of $O_{q\ dEGG}$ a delicate business, which is not so similar with verification of $f_{0\ dEGG}$: it requires more than just a few adjustments for peculiar situations.

PeakDet will ask the verification of $O_{q\ dEGG}$ after the verification of $f_{0\ dEGG}$ has been completed. At first, it will process automatically and offer $O_{q\ dEGG}$ calculated in four different ways:

1. maxima[2] on unsmoothed dEGG signal (displayed as orange squares);
2. maxima on smoothed dEGG signal (displayed as orange stars);
3. barycentre of peak on unsmoothed dEGG signal (displayed as blue squares);
4. barycentre of peak on smoothed dEGG signal (displayed as blue stars)

The methods are divided into two sets:

- Detection of the local minimum on the signal in-between two closure peaks. This method is applied twice: on the unsmoothed dEGG signal, and on the smoothed dEGG

[2]Technically, *'maxima'* here should be referred to as 'minima', since the peak is a negative peak.

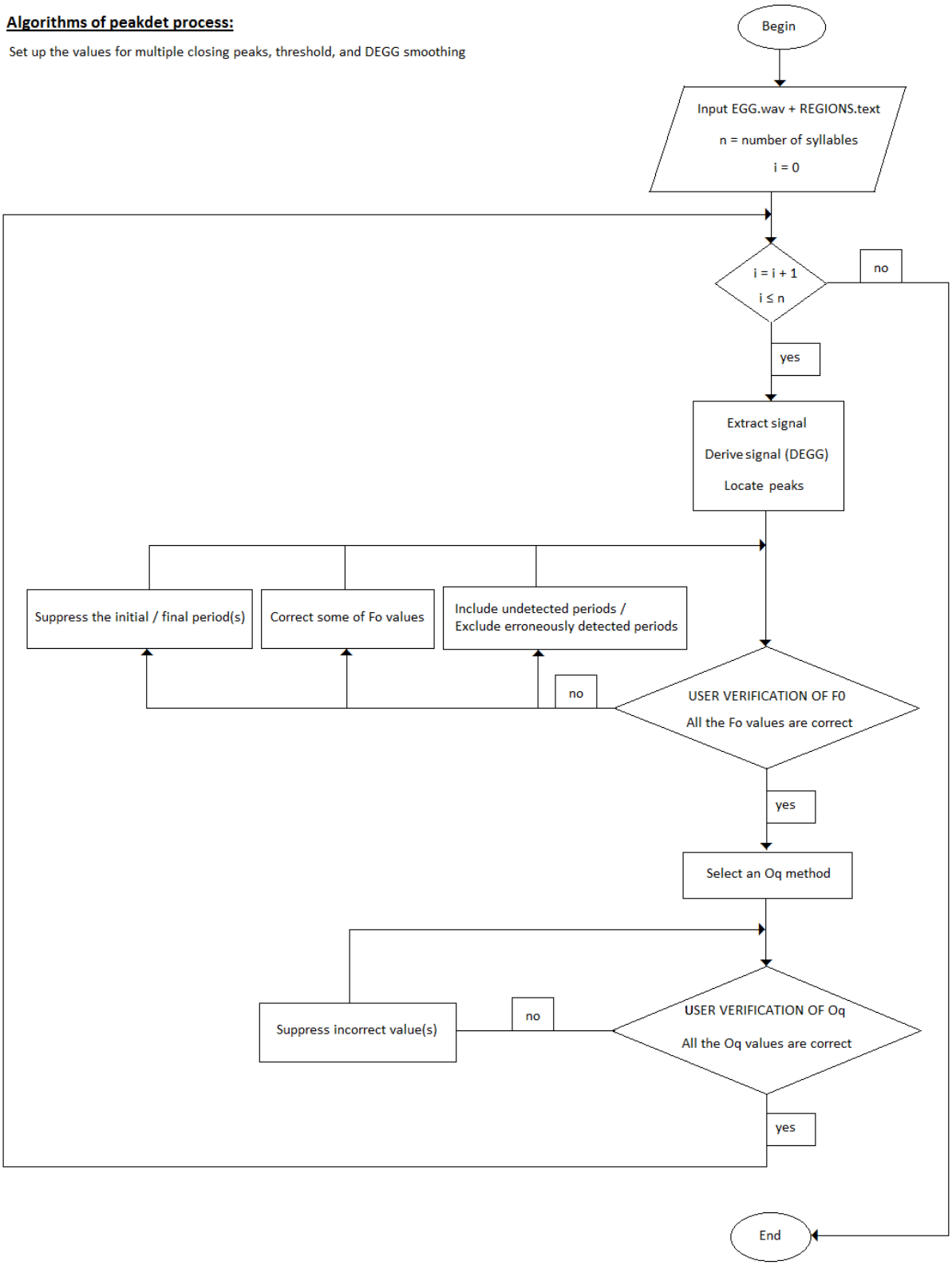
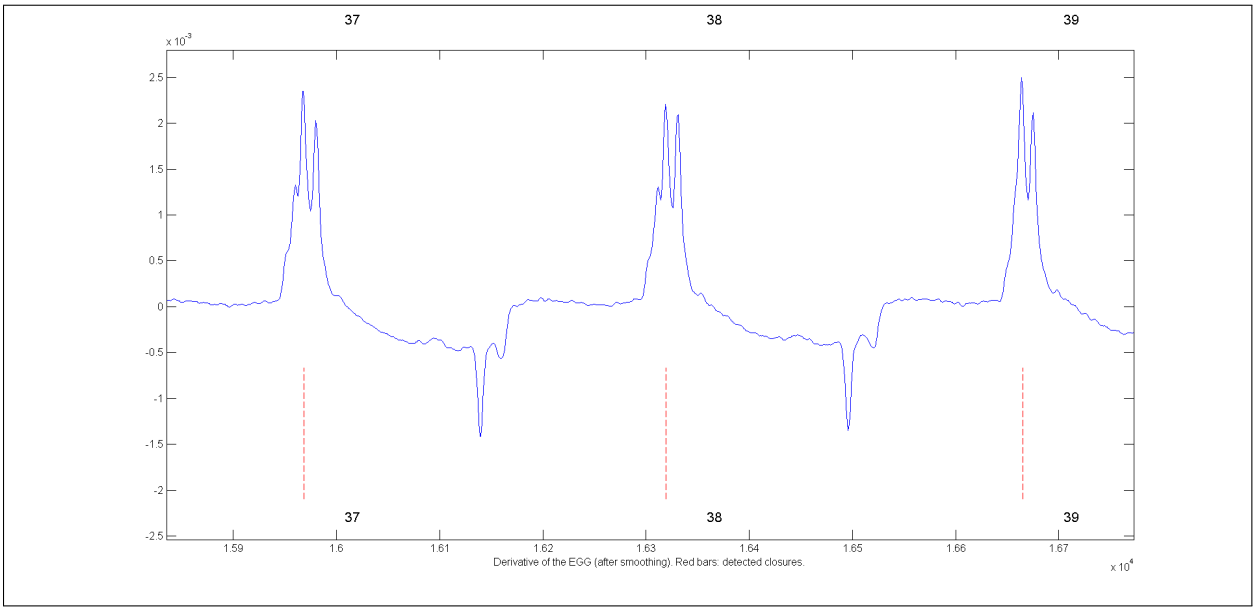
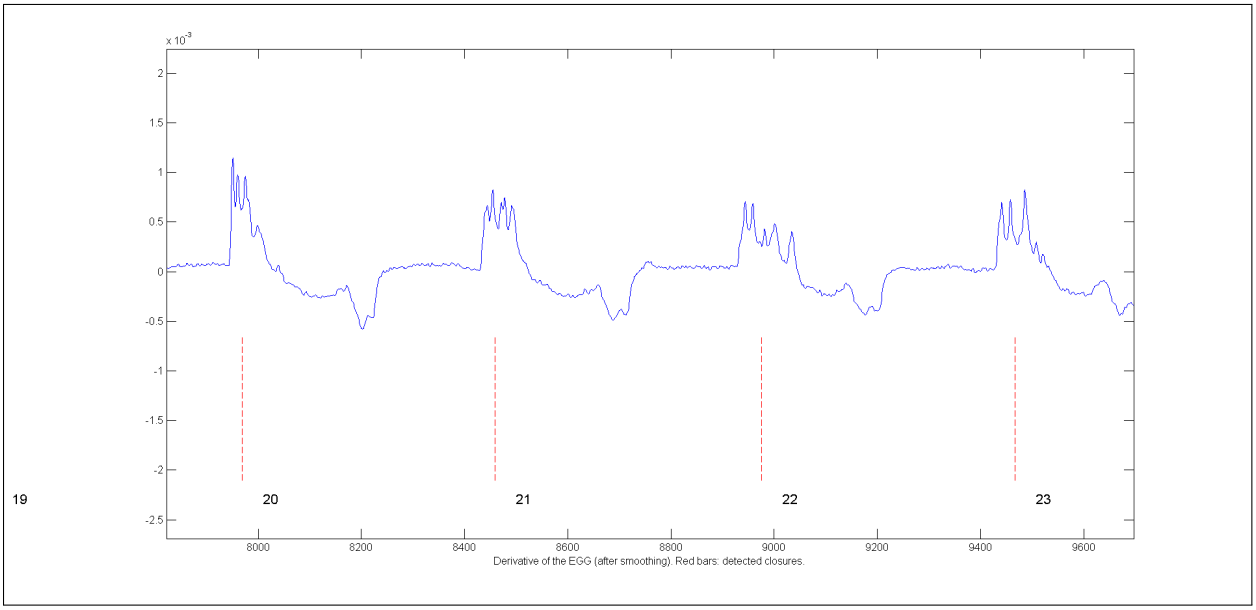


Figure 2.4: A schematic representation of data processing with PeakDet.



(a) Case of double-closing peaks



(b) Case of multi-closing peaks (jagged shape)

Figure 2.5: Examples of imprecise closing peaks. PeakDet uses the method of barycentre automatically to treat these cases. Data from M14.

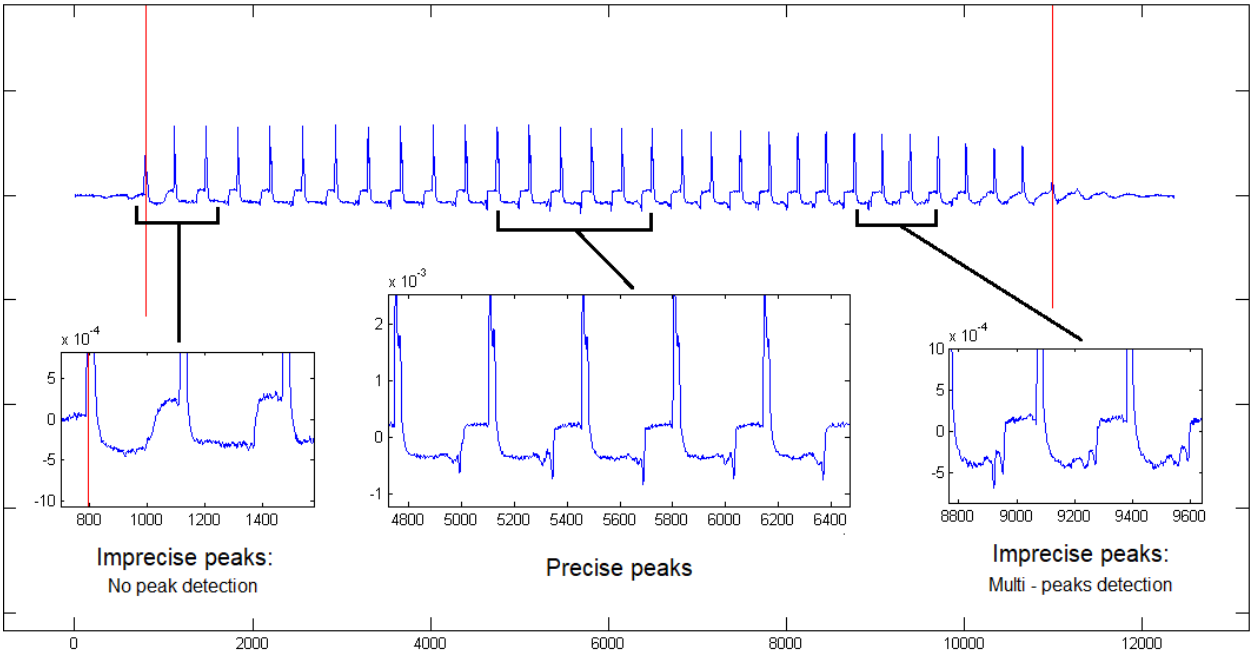


Figure 2.6: Examples of precise and imprecise opening peaks on DEGG signal during the same syllable. Abscissa: in samples (1 sample = 1/44,100 second).

- Analysis of the shape of opening peaks and calculation of a barycentre of the detected 'peaks-within-the-peak', giving each of the peaks a coefficient proportional to its amplitude. Again, this method is applied twice: on the unsmoothed dEGG signal, and on the smoothed dEGG.

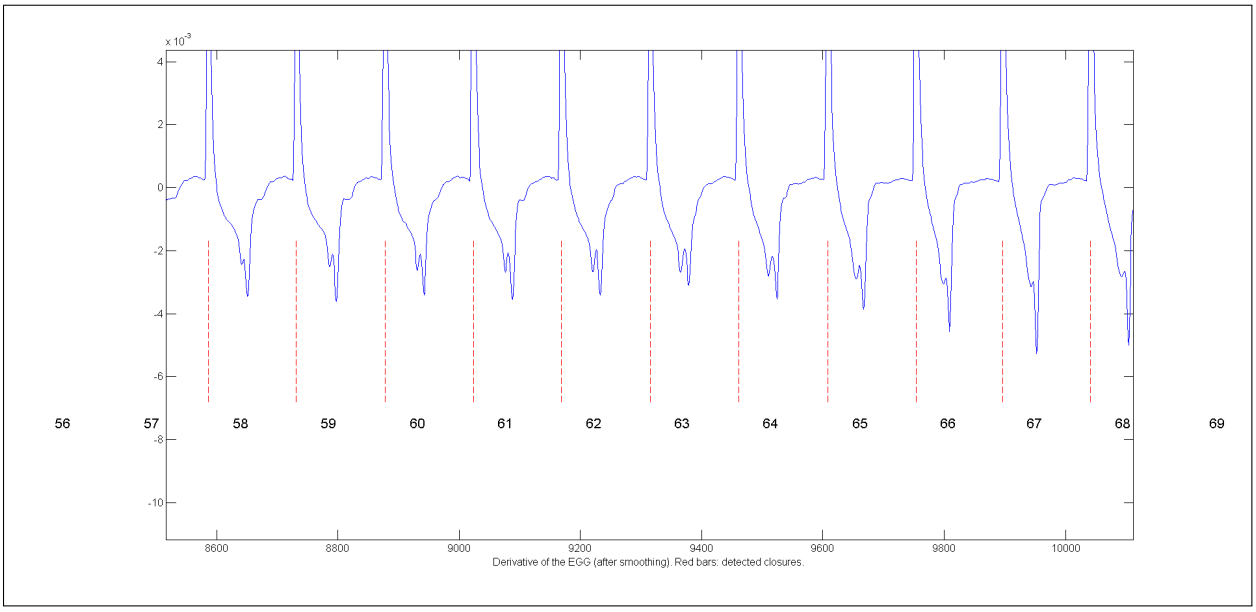
The manual task in this step requires the researcher to visually examine the dEGG signal to decide which method is most reliable for detecting negative peaks, which directly affects the calculation of $O_{q\ dEGG}$ values. This is a two-step verification. The first step is to decide which method will be selected as the most reliable among the four methods, so that the $O_{q\ dEGG}$ values calculated from that method will be stored in the last column (10th column) of the PeakDet results matrix. The second step is to examine each glottal cycle of the dEGG signal in detail to check which cycles have a precise opening instant with a clear negative peak, which will be retained as reasonable $O_{q\ dEGG}$ values for those cycles. Otherwise, cycles that do not have a clear negative peak, because they have multiple peaks or no clear negative peak can be detected (as demonstrated in Figure 2.6), will be removed by setting them to zero to indicate that the automatically calculated $O_{q\ dEGG}$ value for these cycles is inapplicable.

If the decision to retain or remove $O_{q\ dEGG}$ values is simply based on whether or not a single precise opening peak is present in a glottal cycle, it would not be challenging to do it automatically and would not require much effort to visually verify each glottal cycle of each syllable, which is the most time-consuming task, and thus the biggest hurdle for the EGG analysis. However, since information on $O_{q\ dEGG}$ is frequently missing due to the absence of a single, clear opening peak during glottal cycles, it is worth trying to keep the $O_{q\ dEGG}$ values nearly clear in the cases where there is more than one opening peak but visual inspection reveals that one peak really stands out from the others. Two examples in Figure 2.7 show "good" cases of multiple opening peaks: cases where a prominent peak can be noticed in almost every cycle, and the distance between peaks inside the same cycle is small. In cases like this, as a rule I tried to keep as many $O_{q\ dEGG}$ values as possible by choosing the most reasonable method. For example, in the case of Figure 2.7a, I would select the method of barycentre and keep all the $O_{q\ dEGG}$ values because, even though the double-opening peak in cycles 62-64 are clear (compared to its neighboring cycles) and thus make it difficult to decide which one should be the main

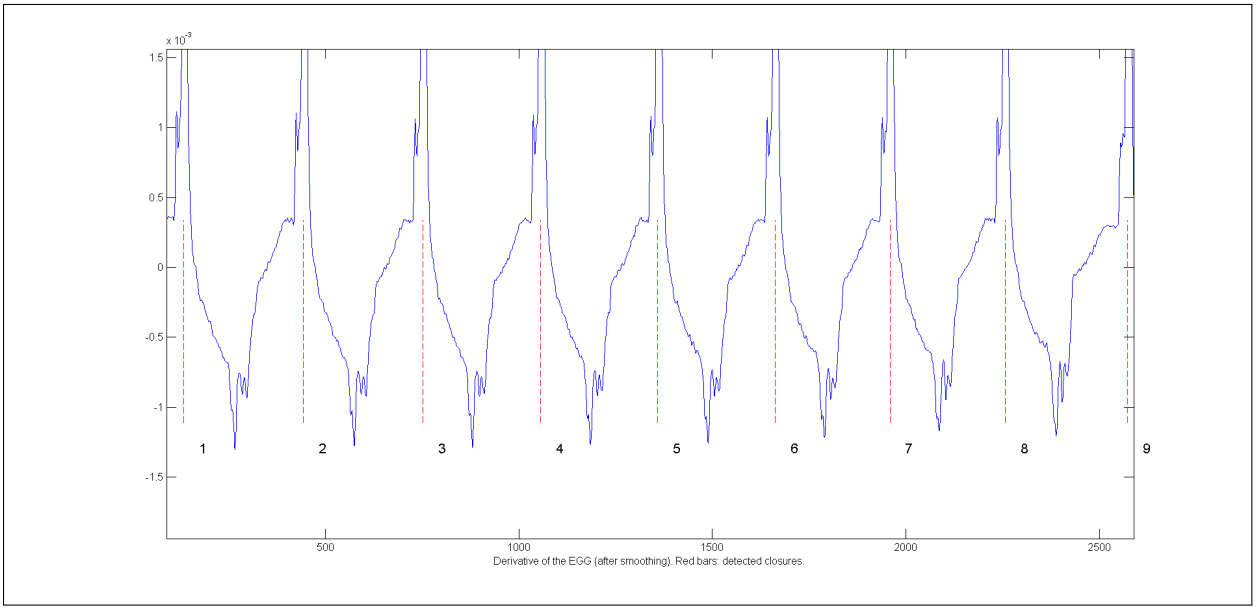
opening instant, the distance between two peaks is nonetheless close enough that it is safe to take an intermediate point between them (by barycentre method) as representative of the approximate opening instant. In the second example in Figure 2.7b, all glottal cycles have triple-opening peaks but the first peaks are always much more salient than the next two peaks. In that case, I would select the method of maxima on dEGG signal to catch the first peak at every cycles as the opening instant used in the calculation of the open quotient.

These two examples illustrate the fact that it is feasible to keep some $O_{q\ dEGG}$ values in case of "good" multi-opening peaks: as long as the main peaks stand out clearly and the distance bettwen them is small (less than 5% difference between methods of measuring $O_{q\ dEGG}$). In practice, there are many tricky cases, where the decision of choosing a method and suppressing certain $O_{q\ dEGG}$ values is much more delicate and tough, particularly when there is a transition between "good" peaks and "hopeless" peaks (Figure 2.7a is a simple example of this). Doing (human) visual verification is not a simply task, and I am not completely confident that my decisions were fully consistent during the analysis (and even if I were confident, *proving* that consistency was present is yet another matter). These observations are crucial to the investigation reported here, as they make it clear that we are totally aware that estimating $O_{q\ dEGG}$ automatically on the basis of an electroglottographic signal is a challenge for neural networks to learn.

In view of the information set out so far, we can now move on to the core of the report: investigating to what extent neural networks can learn from human decisions (as encapsulated in the available corpus) to carry out O_q estimation from the dEGG signal.



(a) Case of "good" double-opening peaks to keep with the method of barycentre of peak



(b) Case of "good" triple-opening peaks to keep with the method of maxima on dEGG signal

Figure 2.7: Examples need manual visual verification. Cases where multi-opening peaks appear but are worth retaining with different methods.

Chapter 3

$O_{q\ dEGG}$ estimation using machine learning methods

This chapter[1] presents a set of experiments whose aim is to select among several estimations of the open quotient for each cycle of an EGG signal. The basis for selection consists of two inputs of a different nature: (i) the EGG signal [2] and (ii) the output of PeakDet, consisting in the time codes of each EGG cycle, 4 candidate values based on the four methods presented in Section 2.3.3, as well as the $f_{0\ dEGG}$ of each cycle also computed by PeakDet. The objective of this section is to describe a neural network that reproduces the manual annotation pipeline described in Section 2.3, i.e. predicting for each cycle whether the cycle has a computable $O_{q\ dEGG}$, and if so, what the most appropriate method is (among the 4 methods).

3.1 Data Description and Preprocessing

The results from manual processing on MATLAB are extracted and stored in excel files (one excel file for each speaker's data) which are made available on a Github repository (see: https://github.com/MinhChauNGUYEN/CLD2025_EGG). Each excel file includes the following information:

- (Column A) The UID (Unique Identifier) of the item.
- (Column B) The beginning time of the syllable/item in the recording (in second)
- (Column C) The end time of the syllable/item in the recording (in second)
- (Column D) The beginning time of each glottal cycle inside the syllable
- (Column E) The end time of each glottal cycle cycle inside the syllable
- (Column F) f_0 values
- (Column G) Oq values as automatically calculated by PeaKDet, method: maxima on unsmoothed signal

[1] Part of this chapter is adapted from Nguyễn, Coavoux, and Rossato (2022).

[2] All data sets of 20 speakers are available on the Pangloss Collection (https://pangloss.cnrs.fr/corpus/Mường). The audio and EGG signal are both freely accessible and downloadable under the CC BY-NC-SA 3.0 license, in the spirit of the *open data* movement in phonetic research (Garellek et al., 2020).

- (Column H) Oq values as automatically calculated by PeaKDet, method: maxima on smoothed signal
- (Column I) Oq values as automatically calculated by PeaKDet, method: barycentre of peaks on unsmoothed signal
- (Column J) Oq values as automatically calculated by PeaKDet, method: barycentre of peaks on smoothed signal
- (Column K) The Oq values that were retained after checking the opening peaks in the DEGG signal (by the user). The zeros mean that the Oq values at these cycles have been suppressed due to the imprecise opening peaks.
- (Column L) The result of a Creak Detection algorithm: (0) means no creak, (1) means pressed voice or single-pulsed creak, (2) means aperiodic creak, and (3) means double-pulsed creak.

Columns B-J will be used as the input to the machine learning system (together with the EGG signal), whereas column K codes the target we need to predict. We present some statistics about the numbers of syllables, glottal cycles and distribution of labels in Table 3.1.

Section	Train	Dev	Test	Complete Corpus
Number of syllabes	9050	1913	2011	12974
Number of glottal cycles	295753	62350	65727	423830
Label distribution				
0 = No $O_{q\ dEGG}$ computable	69237 (23.41%)	14583 (23.39%)	14670 (22.32%)	98490 (23.24%)
1 = maxima without smoothing	60527 (20.47%)	13227 (21.21%)	13461 (20.48%)	87215 (20.58%)
2 = maxima with smoothing	137461 (46.48%)	29909 (47.97%)	30386 (46.23%)	197756 (46.66%)
3 = barycentre without smoothing	42603 (14.40%)	7870 (12.62%)	10102 (15.37%)	60575 (14.29%)
4 = barycentre with smoothing	93583 (31.64%)	18775 (30.11%)	21723 (33.05%)	134081 (31.64%)

Table 3.1: Statistics on the corpus. Since several of the 4 methods implemented in PeakDet may sometimes output the same $O_{q\ dEGG}$ values, some cycles may have several correct labels. As a result the sum of percentages is higher than 100%.

We performed two preprocessing steps, on the wav files in the corpus: (i) segmentation of the signal in multiple shorter files (10 seconds maximum) so that each file contains one or a small number of items/syllables, and (ii) resampling to 32,000 Hz.

3.2 Neural Network Architecture

The machine learning system we implemented is a neural network based on a bi-LSTM. It was implemented using the Speechbrain library (Ravanelli et al., 2021). As stated above, its input consists in the EGG signal, together with additional information for each glottal cycle, namely: time codes (start time and end time), the $f_{0\ dEGG}$ value and the 4 $O_{q\ dEGG}$ values computed by PeakDet. The EGG signal is represented with Mel-frequency Cepstral Coefficients (MFCC) vectors, with a 6 ms window, sliding every 2 ms (these values are lower than values typically used in speech recognition, in order to take into account the granularity of the representations we need).

MFCC vectors form a $N \times F$ matrix $\mathbf{M}^{(o)}$, where N is the number of MFCC frames (i.e. the length of the signal in milliseconds divided by 2) and F is the number of MFCC features for each frame. This

matrix is fed to a feedforward neural network:

$$\mathbf{M}^{(1)} = \tanh(\mathbf{W}^{(1)} \cdot \text{LayerNorm}(\mathbf{M}^{(0)}) + \mathbf{b}^{(1)}),$$

and contextualized with a bidirectional LSTM:

$$\mathbf{M}^{(2)} = \text{bi-LSTM}(\mathbf{M}^{(1)}). \tag{3.1}$$

Then, we represent each glottal cycle c by the concatenation of 3 vectors:

$$\mathbf{v}_c = [\mathbf{M}^{(2)}_{c_b}; \mathbf{M}^{(2)}_{c_e}; \mathbf{o}_c],$$

where c_d and c_f are the time codes for the (b)eginning and (e)nd of the cycle, and $\mathbf{v}_c \in \mathbb{R}^5$ is a vector containing the 4 $O_{q\ dEGG}$ values and the $f_{0\ dEGG}$ of the cycle, as computed by PeakDet. Finally, we use another feedforward network to compute scores for each label and predict a label:

$$\mathbf{P} = \text{Sigmoid}(\mathbf{W}^{(3)} \cdot \text{ReLU}(\mathbf{W}^{(2)} \cdot \text{LayerNorm}(\mathbf{v}_c) + \mathbf{b}^{(2)}) + \mathbf{b}^{(3)}),$$

where each $\mathbf{P} = [P(y_0 = 1|c), \ldots, P(y_4 = 1|c)]$ gives a probability for each label. An illustration of the neural network is presented in Figure 3.1.

3.3 Experiments

Training We train the model by maximizing the probabilities of gold labels on the training set, using the Adam algorithm (Kingma and Ba, 2015). When evaluating the model, we take the highest probability label as the model's prediction. The hyperparameters of the model are:

- For MFCCs: window size (6ms), hop size (2ms), context size (2 frames on each side);
- For the network: dimension of hidden layers (128 for feed-forward networks, 128 for each direction of the bi-LSTM);
- For training: optimization algorithm (Adam), learning rate (0.008 for models that use the signal as input, 0.001 for the model ablation that does not use the signal), size of batches (8), number of training epochs.

We calibrated hyperparameters (in particular the learning rate) on the dev set during preliminary experiments. For final experiments, we train the models for 100 epochs and keep the checkpoint that maximizes accuracy on the dev corpus to evaluate it on the test section.

Experimental settings Our objective is to determine whether the use of the EGG signal improves the prediction of labels, and contributes additional information compared to PeakDet $O_{q\ dEGG}$ and $f_{0\ dEGG}$ values. Moreover, we would like to assess whether the EGG signal provides better information than the more easily accessible audio signal. To answer these questions, we experiment with several configurations, namely models with different types of input:

i EGG signal + PeakDet $O_{q\ dEGG}$ + PeakDet $f_{0\ dEGG}$: $[\mathbf{M}^{(2)}_{c_d}; \mathbf{M}^{(2)}_{c_f}; \mathbf{o}_c]$;

ii EGG signal: $[\mathbf{M}^{(2)}_{c_d}; \mathbf{M}^{(2)}_{c_f}]$;

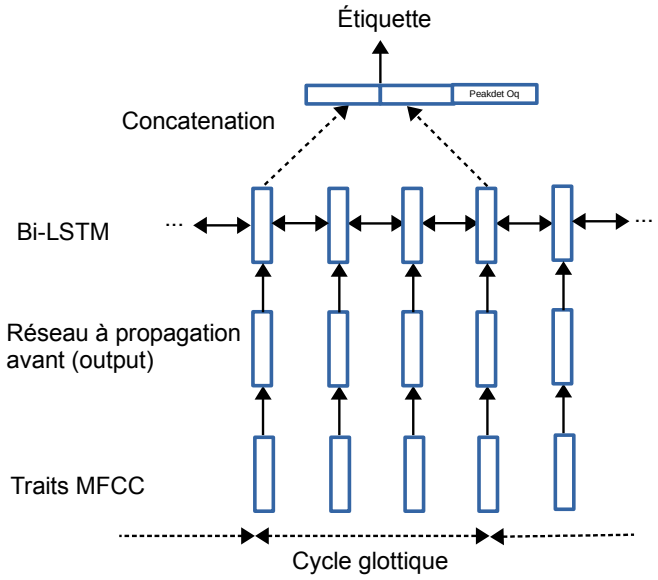


Figure 3.1: Neural network architecture (centered on a single glottal cycle) In practice, the bi-LSTM encodes a full syllable.

iii PeakDet $O_{q\ dEGG}$ + PeakDet $f_{0\ dEGG}$: $[\mathbf{o}_c]$;

iv as (i) but replace EGG signal by audio signal;

v as (ii) but replace EGG signal by audio signal.

3.4 Results and discussion

We report results in Table 3.2. We use the following evaluation metrics: 3-class accuracy (0 vs 1-2 vs 3-4), 2-class accuracy and 2-class Fscore (label 0 vs other labels). These metrics are meant to take into account class imbalance as well as massive overlap between respectively labels 1-2 and labels 3-4. For comparison, we also present results for a most-frequent-label baseline.

From Table 3.2 we see that the neural network outperforms the baseline, albeit by a small margin, indicating that some learning has occurred but modestly. Model (ii), with only the EGG signal as input, has lower results than model (iii), which suggests that the shape of the signal as provided as input is either insufficient to make a prediction, or underexploited by the model. The best models (i, iv) have access to both a signal and the additional information (PeakDet $O_{q\ dEGG}$ and $f_{0\ dEGG}$). Unexpectedly, models (i) and (iv) have comparable scores, with a slight advantage for using the audio signal (iv) instead of the EGG signal (i).

This conclusion is indeed unexpected and not good news for those who, like us, work on the EGG signal, but it is actually not totally bizarre and unreasonable. If we look back and compare all the available signals of syllables bearing a creaky tone, in many cases such as item in Figure 3.3 is an example, we could notice that, in fact, the audio signal with its spectrogramme, already provide the good evidence for the presence of creaky voice, even much clearer visually in compararison to the EGG signal with the series of discrete rails along the whole syllable.

	Dev			Test		
Model	Accuracy-3	Accuracy-2	Fscore-2	Accuracy-3	Accuracy-2	Fscore-2
Baseline (most-frequent class)	48.9	76.6	0	46.2	77.7	0
(i) EGG + O_q + f_0	**63.6**	**86.1**	**91.0**	58.4	**85.8**	**91.0**
(ii) EGG	56.9	80.6	87.7	53.2	78.2	86.1
(iii) O_q + f_0	58.7	83.2	89.1	56.9	83.6	89.5
(iv) Audio + O_q + f_0	63.4	85.9	**91.0**	**59.4**	85.2	90.6
(v) Audio	57.1	78.9	86.6	51.5	79.2	87.3

Table 3.2: Final results on development and test sets (%).

pred ↓ gold→	0	1	1/2	2	2/3	2/3/4	2/4	3/4	4
0	61.77*	6.9	6.46	9.24	4.11	1.1	1.61	3.27	13.13
1	0	0*	0*	0.01	0	0	0	0	0
2	26.79	89.66	73.2*	72.67*	76.71*	74.88*	79.03*	60.23	50.35
3	0.04	0	0.29	0.22	0*	0.37*	0	0.46*	0.08
4	11.4	3.45	20.05	17.87	19.18	23.66*	19.35*	36.04*	36.44*

Table 3.3: Confusion matrix (test corpus) for model (iv). Values are percentages computed on gold labels. Correct predictions are indicated with a star (*).

Nevertheless, the negative results obtained here by machine learning do not mean that the phonetic analysis of the EGG approach should be rejected or discredited but instead it brings us back to the fundamental issue on the reality of these signals. Ultimately, neither of these signals is fully representative of what actually happens during speech articulation. They are all linear views of a non-linear phenomenon. They are therefore not in a mutually contradictory or mutually exclusive relationship, but can actually support each other in the analysis of complex phonetic characteristics. We chose to collect and analyze the EGG signal not because we think it is better than the acoustic signal, but in the hope that it might provide an additional/alternative approach to the unresolved limitations of audio signals on speech analysis. In this corpus, all the data of EGG signal were recorded and stored simultaneously with the acoustic signal. This allows us, after taking a big jump from manual processing to end-to-end processing, to step back and see the possibility of improving the current task, for instance by further examining the correlation between EGG and acoustic signals on the same parameters to figure out how they could complement and reinforce each other in the automatic workflow.

Confusion matrix Going into details of the results, we present in Table 3.3 the confusion matrix of model (iv), which obtained the best results. Each column represents a combination of labels seen in the data (e.g., column 3/4 represents glottic cycles where PeakDet methods 3 and 4 gave the correct value.) We observe that labels 1 and 3, which are the least present in the data (and often give the same result as methods 2 and 4 respectively), are almost never predicted, which shows that the system does not discriminate between classes 1 and 2 on the one hand and classes 3 and 4 on the other hand. The most frequent errors are the difficulty in predicting classes 3, 4 and 0, where the model falls back on the most frequent class (2). This reflects to some extent what actually happened during the semi-automatic processing.

As mentioned in Section 2.3.3, among 4 proposed options of $O_{q\ dEGG}$ verification on PeakDet, actually there are only two different methods which is: (i) "maxima" (detection of the local minimum on the signal in-between two closure peaks) in methods 1 and 2 and (ii) and "barycenter" (analysis of the shape

of opening peaks and calculation of a barycentre of the detected 'peaks-within-the-peak', giving each of the peaks a coefficient proportional to its amplitude.) in methods 3 and 4. Then, each method is subdivided according to whether it was applied to the unsmoothed dEGG signal (method 1 or 3) or the smoothed signal (method 2 or 4).

Since the openning peak of a glottic cycle is less clear than the closing peak due to one of the facts that multiple negative peaks can often be detected during the open phase, methods applying on the smoothed signal are therefore preferable to unsmoothed methods in order to avoid redundant detection of peaks caused by signal noise. This explains why methods 1 and 3 are much less selected than methods 2 and 4. Even in cases where there is a clear opening peak for which all four methods give identical or quasi-identical $O_{q\ dEGG}$ values, I kept selecting the methods on the smooth signal to keep the results consistent.

The option of barycenter method on unsmoothed dEGG signal (method 3) was barely selected, as it is the least appropriate in all cases, whether it is a single peak or a multiple peak. Cases like in the examples 2.7a or 3.2 that involve double-opening peaks or multiple-opening peaks but the distance is safely closed than thus it is worth keeping some of them with the barycenter method. And the one applied to the smoothed signal (i.e., method 2) is the most reliable option to eliminate all negative peaks that are not prominent enough. This explains why method 2 has been the most frequent option, as it is the safest in most cases, especially for "good" multiple-opening peaks.

On the other hand, the local minimum method on the unsmoothed signal (method 1) was particularly used in a few cases involving creaky voice. During manual processing, we observed that in most cases, once creaky voice occurs, it frequently causes irregularities in the glottal cycles. The mess in the EGG signal and consequently in the dEGG signal in this case does not imply a bad signal due to a recording artifact, but in fact, it faithfully reflects a messy vocal fold contact area. The loss of periodicity is one of the main factors that make the analysis of creaky voice in particular and glottalization in general a real challenge. Every single value is precious for assessing the phonetic characteristics of this non-modal vocal quality. Therefore, the primary goal in the manual analysis of the data was to try to retain as much information as possible about the values in the tokens containing the creaky voice, without of course falling into the "creaky voice lover" bias. This is to say that in this case (where the barycenter method is hopeless), it makes sense to consier picking either method 1 (minimum in the unsmoothed signal) or method 2 (minimum in the smoothed signal) to ensure that the most reliable values will be retained, while undergoing an honest verification. We know that creaky voice is the lowest voice quality with low f_0 and O_q. But the choice of method 1 or 2 was not driven by having the lowest recorded values, but by having the most reliable values of the creaky portion where each cycle is visually verified and ensured that there is a prominent negative peak worth saving.

Figure 3.3 provides an example of a syllable bearing a creaky tone (data from speaker F3, syllable /na4/ "archery") [3]. As the principle just mentioned, in this case, since most glottal cycles have multiple opening peaks along the open phases, two methods of barrycenter are appropriate. The consideration here will be only the two methods of local minimum peak on the dEGG signal, i.e. method 1 or 2. Despite the messy glottal cycles, we still can clearly spot that at many cycles a good negative peak could be found among plenty of minor peaks. In a zoom on cycles 6 to 10, cycles 6 - 7 - 9 clearly have a precise opening peak, thus should be retained, whereas cycles 8 and 10 also tend to have a major peak at 2/3 portion of the open phase but much less clear, thus should be eliminated. In general, for the whole syllable, method 1 often has lower $O_{q\ dEGG}$ values than method 2, but in this case I chose method 2 over method 1 because it gives a correct value more on the cycle 6th with a good opening peak as verified in the previous step. The analysis and examination was carried out with such a delicate and meticulous process. It intrigued me how the neural network could learn and perform the same task.

[3] The data of this example is available here: https://doi.org/10.24397/pangloss-0006761#W90

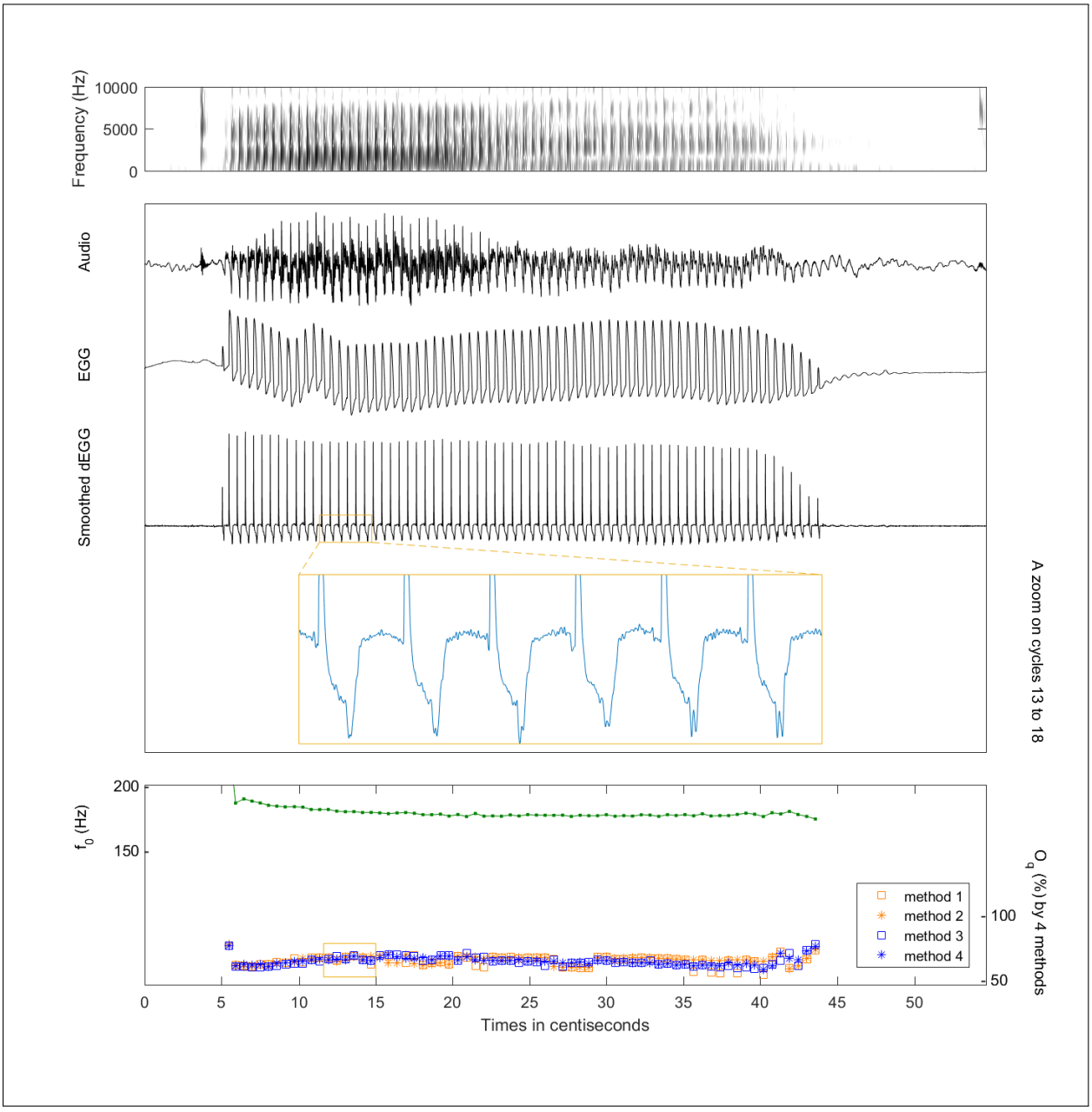


Figure 3.2: $O_{q\ dEGG}$ verification: case where methods 3 and 4 (barycenter of the peaks on the dEGG signal) have been considered to be chosen.

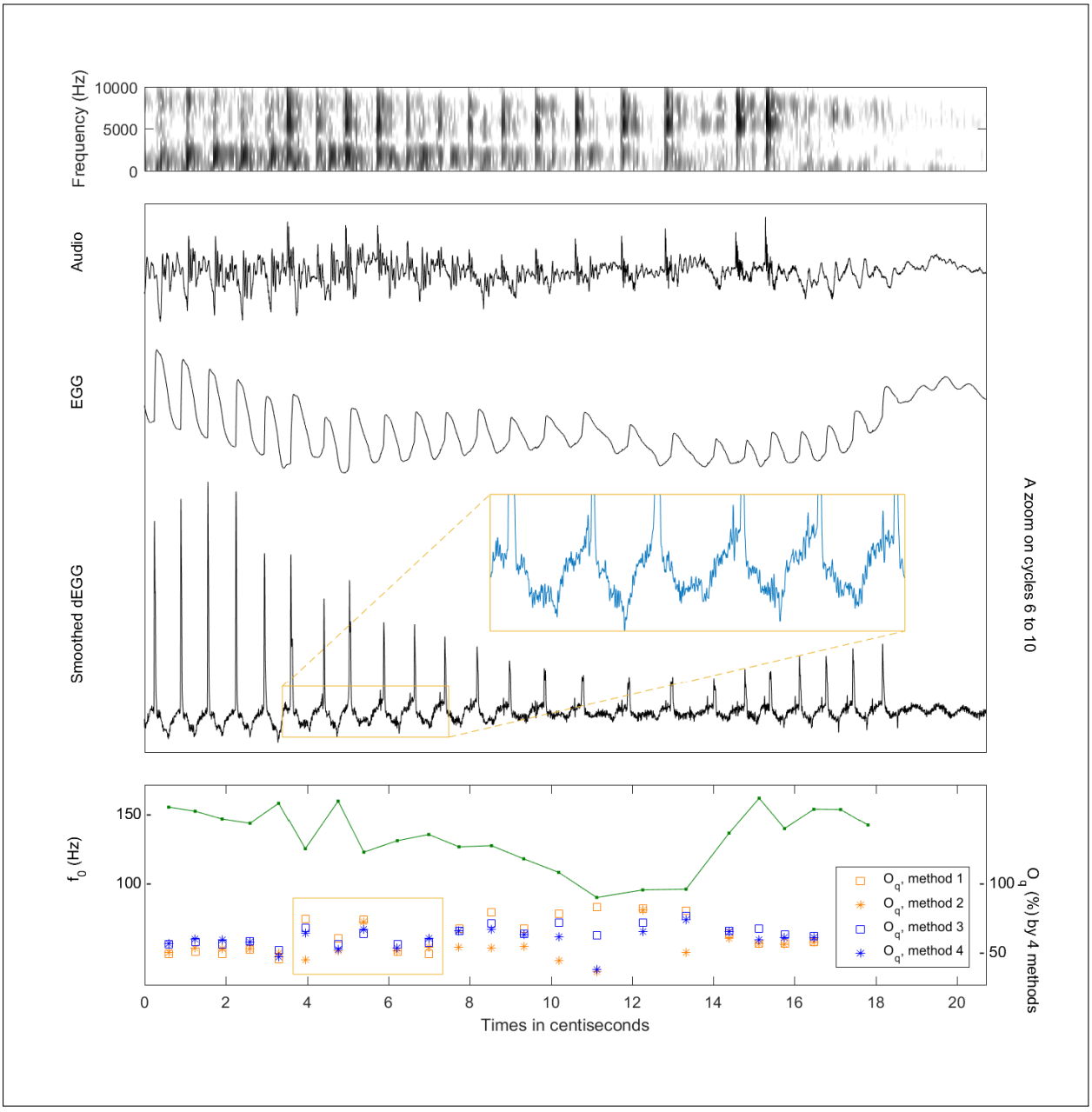


Figure 3.3: $O_{q\ dEGG}$ verification (cont.): case where methods 1 and 2 (local minimum on the dEGG signal) have been considered to be chosen.

Limitations In retrospect, it becomes clear that the task is too hard to be addressed with an end-to-end statistical tool, used "off-the-shelf", as it were. Specifically, we identified two weaknesses in the design of the experiments we have presented.

First, the data includes both target syllables and frame syllables from carrier sentences. Since carrier sentences remain stable (quasi identical) across items, the model might be biased towards easier predictions. For a closer look at this issue, we need to refer to the total number of each component of the copus, as summarized in Figure 2.2. In the total number of 660 items for each speaker's data, the number of processed items is 264 in the target syllables and 396 in the frame syllables. Each target syllable was repeated twice in isolation and twice in a carrier sentence (2.1). The frame syllables of the carrier sentence are composed of only three different words that are repeated to carry each of the 132 target syllables, i.e. each frame word will also be repeated 132 times. It is therefore obvious that this data is highly repetitive and simple. The most diverse part is that the target words also belong to 12 minimal sets and 3 minimal pairs, i.e., within the sets, they are completely or almost exactly the same in terms of syllables, and differ only in tone. This is apparently not an ideal corpus for training in automatic machine translation or transcription tasks. Indeed, this corpus was one of the test subjects for the Persephone tool within the framework of the project of phonemic transcription of low-resource languages carried out by Wisniewski, Michaud, and Guillaume (2020) and, as anticipated, with only one hour of training, the results were rather good. In this pilot study exploring the capability of machine learning perspertives for the analysis of electroglottographic signals, despite the simplicity of the corpus, we still obtained a negative result. That shows how challenging this task is.

If the first limitation comes from the inherent design of the data that cannot be changed because the starting point of the study is not initially set for an end-to-end approach, the second limitation comes from the flaw in communication between, on the one hand, the linguists who own and understand the data and, on the other hand, the scientist who takes over the data for training in the neural network model. In detail, the manual annotation process makes 2 decisions: (i) a decision on the best PeakDet method (at the level of the syllable) (ii) a decision about whether to keep or discard the $O_{q\ dEGG}$ (a decision made at the level of glottal cycle). In contrast, our model only makes predictions at the level of the cycle, thus making the task harder. Aggregating cycle-level predictions into a syllable-level prediction (e.g. through voting) might lead to better results. This limitation is due to a misunderstanding about the process of verifying $O_{q\ dEGG}$ using PeakDet. The most effective way to explain this would be to make a demo of what actually happened during the semi-automatic processing using PeakDet. However, due to a technical problem (PeakDet was not working properly on my current computer), I was not able to do this demo directly, and apparently written explanations, even if very specific and detailed, are still not sufficient for data scientists to receive and understand the data thoroughly. As a consequence, the process of machine learning (learning patterns of statistical association between data and labels) clearly appears not to have followed the path that I (somewhat naively) expected, namely: kindly following in my footsteps, emulating the analytic process that I had adopted.

Such a finding is by no means new: features extracted by end-to-end models should not be expected to match the features commonly used in 'manual' workflows. Thus, Gendrot, Ferragne, and Chanclu (2022) report that a convolutional neural network that is (moderately) successful at speaker classification makes use of spectral and temporal features that are not related to classical phonetic measures in any straightforward way. Such observations open into an exciting mid-term research program: studying statistical models to see how they encapsulate relevant information, and how this information can shed light on the languages found in the datasets used at training. No more will be said on this topic here, though, as studies on explicability of neural models are best based on highly successful models, whereas the experiments reported here did not yet reach the level of practical usefulness (which would be technically reflected in low error rates).

Study of phonation types The multidisciplinary work described here has implications for the phonetic study of phonation types. Indeed, the set of results obtained here highlights the fact that the process of assessing the reliability of the open quotient estimate in view of the signal is a non-trivial task. This provides an opportunity to return to the fundamental question of what the glottic open quotient reflects, and how it is interpreted. The open quotient is a linear projection of the non-linear phenomena involved in phonation, and therefore obviously cannot by itself be a sufficient descriptor of the various types of phonation: whispered voice, pressed voice, cracked voice; among the commonly used references, see in particular Laver (1980). Specifically in the case of creaky voice, we observe a good correlation between the presence of creaky voice and the spectral slope information reflected in the spectrogram (a higher intensity in the upper half of the spectrogram, from 5 to 10 kHz, than in the lower half). By contrast, the glottal open quotient does not clearly show the same degree of correlation with creaky voice. Thus, the audio signal can sometimes be a better guide than the open quotient for detecting creaky voice. The EGG signal contains other information – most obviously the fundamental frequency – which provides clearer indications than O_q regarding the phonatory type.

Acoustically, it is known that the open quotient is neither the only nor the most important of the glottal source parameters. The fact that it can be estimated from the EGG signal has undoubtedly led to it being given particular importance, in comparison, for example, with the speed quotient, which is not easily accessible for estimation. Thus, the height of the positive peak on the derivative (corresponding to the moment of glottic closure at the beginning and end of the cycle), DECPA (for Derivative-Electroglottographic Closure Peak Amplitude: Michaud 2004a), does not constitute a means to estimate the speed quotient robustly and reliably. Clearly, O_q would benefit from being integrated into machine learning experiments in which it would be integrated into a larger set of acoustic parameters, in order to characterize various types of phonation in an objective and complete way.

Perspectives In future work, we consider developing and evaluating regression models that directly predict $O_{q\ dEGG}$ as a continuous variable, instead of predicting it indirectly through a classification task. Another future direction that we consider investigating consists in the replacement of MFCC features by available multilingual pretrained acoustic models (Conneau et al., 2020), which have rapidly become mainstream in speech processing (including applications to fieldwork corpora: Guillaume et al. 2022).

Appendices

Appendix A

Detailed information on the speech material and the total corpus

A.1 Speech material

Tables A.1 and A.2 provide full detail about the minimal sets and pairs. The tables include:

- First column: The numbering of minimal sets (from 1 to 12) and minimal pairs (from 1 to 3).
- Second column: The numbering of target syllables, labeled as "UID" (for "Unique Identifier") because this number constitutes the unique identifier of target syllables. This number is used in the annotation of audio files, and in data processing down the line.
- Third column: The target syllables. These constitute the actual speech material of the recording session. In other words, the speakers were asked to pronounce these monosyllabic morphemes (roots).
- Fourth column: The full form of the target words from which monosyllables were extracted, in cases where the usual form of the word at issue is disyllabic. This point will be elaborated on below.
- Fifth-sixth-seventh columns: The translations in English, French and Vietnamese, respectively.

This information is important to know in order to be able to retrieve the content of the data later, as it is encoded as labels in the annotation list.

Table A.1: Speech materials: eight minimal sets and four near-minimal sets that contrast for five tones in smooth syllables.

N.	UID	Target syllable	Complete word	English	French	Vietnamese
1	1	**paj^{5}**	**paj^{5} t^{h}ăj^{1}**	arm span	empan (de bras)	sải tay
	2	**paj^{3}**	**ke^{4} paj^{3}**	a cylindrical jar to ferment vegetables	un pot cylindrique pour la fermentation des légumes	vại
	3	**paj^{2}**	**tɤ̆p^{6} paj^{2}**	barrage	barrage	đập tràn
	4	**paj^{1}**	**paj^{1}**	cloth	tissu	vải
	5	**paj^{4}**	**paj^{4}**	fruit	fruit	quả
2	6	**rɔ^{5}**	**kɔn^{2} rɔ^{5}**	tortoise	tortue	rùa
	7	**rɔ^{3}**	**rɔ^{3} kuə^{2}**	to find crab (by hand) in rice field	attraper des crabes (à la main) dans une rizière	mò
	8	**rɔ^{2}**	**rɔ^{2}**	to be sated	être rassasié	no
	9	**rɔ^{1}**	**ʔɤ^{5} rɔ^{1}**	idle	désœuvré	rảnh rỗi
	10	**rɔ^{4}**	**paj^{4} rɔ^{4}**	banana flower	fleur de bananier	hoa chuối
3	11	**pa^{5}**	**pa^{5}**	grand-mother	grand-mère	bà
	12	**pa^{3}**	**pa^{3}**	to touch on one’ s shoulder	se toucher l’épaule	bấu vai
	13	**pa^{2}**	**pa^{2}**	three	trois	ba
	14	**pa^{1}**	**pa^{1} t^{h}iən^{5}**	to pay	payer	trả tiền
	15	**pa^{4}**	**pa^{4}**	to patch	rapiécer	vá (xăm)
4	16	**laj^{5}**	**laj^{5}**	tongue	langue	luỡi
	17	**laj^{3}**	**pɤ^{2} laj^{3}**	to return	revenir	trở lại
	18	**laj^{2}**	**laj^{2}**	carry stuff or people on motorcycle	transporter des objets ou des personnes à moto	lai
	19	**laj^{1}**	**laj^{1} tɤ̆m^{5}**	a bamboo fence to keep fish in the lake	une barrière de bambou pour garder les poissons dans le lac	cái rào ao
	20	**laj^{4}**	**laj^{4}**	to drive	conduire	lái

5	21	**taj^{5}**	**taj^{5} kaw^{4}**	to wash (rice)	laver (riz)	đãi gạo
	22	**taj^{3}**	**taj^{3} tɔŋ^{4}**	to pull bamboo by hand or by motorbike	tirer le bambou à la main ou à moto	lôi bương
	23	**taj^{2}**	**taj^{2} nan^{3}**	accident	accident	tai nạn
	24	**taj^{1}**	**taj^{1}**	cascade	cascade	thác nước
	25	**taj^{4}**	**taj^{4}**	urinate	uriner	đái
6	26	**kɔ^{5}**	**kɔn^{2} kɔ^{5}**	heron	héron	con cò
	27	**kɔ^{3}**	**kɔ^{3}**	to speak	parler	nói
	28	**kɔ^{2}**	**kiəw^{4} kɔ^{2}**	tug of war	lutte acharnée	kéo co
	29	**kɔ^{1}**	**kɔ^{1}**	grass	herbe	cỏ
	30	**kɔ^{4}**	**kɔ^{4}**	to have	avoir	có
7	31	**kiɛŋ^{5}**	**kiɛŋ^{5}**	a earthenware jar to store liquids	une jarre en faïence pour conserver des liquides	bình sứ
	32	**kiɛŋ^{3}**	**ʔʁ^{2} kiɛŋ^{3}**	beside	à côté	ở cạnh
	33	**kiɛŋ^{2}**	**kiɛŋ^{2}**	soup	soupe	canh
	34	**kiɛŋ^{1}**	**kiɛŋ^{1}**	gong	gong	kiẻng
	35	**kiɛŋ^{4}**	**kiɛŋ^{4}**	wing	aile	cánh
8	36	**ma^{5}**	**ma^{5} lɯəɲ^{2}**	ell's cave hole	trou de l'anguille	lỗ lươn
	37	**ma^{3}**	**ma^{3}**	rice seedings	plants de repiquage	mạ
	38	**ma^{2}**	**ma^{2}**	ghost	fantôme	ma
	39	**ma^{1}**	**ma^{1}**	tomb	tombeau	mả
	40	**ma^{4}**	**ma^{4}**	cheek	joue	má
9	41	**ŋa^{5}**	**ŋa^{5}**	to fall	tomber	ngã
	42	**ŋa^{3}**	**ŋa^{3}**	itch	prurit	ngứa
	43	**ŋa^{2}**	**ŋa^{2} măt^{6}**	dazzle	éblouissement	chói
	44	**ŋa^{1}**	**ŋa^{1} lɯŋ^{2}**	to recline	s'incliner	ngả lưng
	45	**na^{4}**	**ke^{4} na^{4}**	archery	archerie	cung tên
10	46	**ka^{5}**	**paj^{4} ka^{5}**	eggplant	aubergine	cà
	47	**ta^{3}**	**ta^{3}**	dumbbell	haltère	quả tạ
	48	**ka^{2}**	**kɔn^{2} ka^{2}**	chicken	poulet	gà
	49	**ka^{1}**	**ka^{1}**	big	grand	to
	50	**ka^{4}**	**ka^{4}**	fish	poisson	cá

11	51	**kaj^{5}**	**kaj^{5}**	to button	boutonner	cài (cúc)	
	52	**paj^{3}**	**paj^{3}**	a cylindrical jar to ferment vegetables	un pot cylindrique pour la fermentation des légumes	vại	
	53	**kaj^{2}**	**kaj^{2}**	thorn	épine	gai	
	54	**kaj^{1}**	**ta^{6} kaj^{1}**	cabbage	chou	(rau) cải	
	55	**kaj^{4}**	**kɔn^{2} kaj^{4}**	the female	la femelle	con cái	
12	56	**ku^{5}**	**ku^{5}**	old	ancien	cũ	
	57	**tu^{3}**	**tu^{3} mɤ̌w^{4}**	hematoma	hématome	tụ máu	
	58	**ku^{2}**	**ku^{2}**	buffalo	buffle	trâu	
	59	**ku^{1}**	**ku^{1}**	tuber	tubercules	củ	
	60	**ku^{4}**	**ku^{4} miɛw^{5}**	owl	hibou	cú mèo	

Table A.2: Speech materials: three minimal pairs that contrast the two tones of checked syllables.

N.	UID	Target syllable	Complete word	English	French	Vietnamese
1	61	**pat^{6}**	**pat^{6} ɲa^{5}**	floor made by a kind of bamboo	plancher fait d'une sorte de bambou	sàn nhà sàn
	62	**pat^{7}**	**pat^{7}**	bowl	bol	bát
2	63	**rwɛc^{6}**	**rwɛc^{6}**	intestine	intestins	ruột
	64	**rwɛc^{7}**	**rwɛc^{7}**	to pour	verser	rót
3	65	**lak^{6}**	**lak^{6}**	peanut	cacahuète	lạc
	66	**lak^{7}**	**lak^{7}**	squint eye	strabisme	lác

A.2 Corpus status

Table A.3: Current status of corpus: 20/28 data files have been annotated with Sound Forge and processed with Matlab.

N°	Speaker	Quality of EGG signal	Data status	Size of .mat file
1	F1	Crackling noise	No annotation	No analysis
2	F1	Crackling noise	No annotation	No analysis
3	F3	Good	660/660 items	100×10×660
4	F7	OK	660/660 items	119×10×660
5	F8	Weak EGG	No annotation	No analysis
6	F9	Good	660/660 items	119×10×660
7	F10	Good	585/660 items Missing 75 items - 11 target words in isolation - 16 target words in carrier sentence - 16 frame words at 1st position - 16 frame words at 3rd position - 16 frame words at 4th position	178×10×585
8	F11	Weak EGG	No annotation	No analysis
9	F12	Good	660/660 items	111×10×660
10	F13	Good	646/660 items Missing 14 frame words at 1st position	133×10×646
11	F14	OK but there are a few flat segments	No annotation	No analysis
12	F16	Weak EGG	No annotation	No analysis
13	F17	Good	660/660 items	100×10×660
14	F18	Weak EGG	No annotation	No analysis
15	F19	Good	660/660 items	100×10×660
16	F20	OK	660/660 items	100×10×660
17	F21	OK	660/660 items	111×10×660
18	M1	Good	660/660 items	100×10×660
19	M5	OK	660/660 items	100×10×660
20	M7	Good	660/660 items	100×10×660
21	M8	OK	660/660 items	100×10×660
22	M9	Good	660/660 items	100×10×660
23	M10	Good	660/660 items	100×10×660
24	M11	Good	660/660 items	100×10×660
25	M12	Signal out of range	No annotation	No analysis
26	M12	Good	660/660 items	100×10×660
27	M13	Good	660/660 items	100×10×660

28	M14	OK	656/660 items Missing 4 frame words: 3 at first position and 1 at 4th position	100×10×656

Bibliography

Abramson, Arthur S, Mark K Tiede, and Theraphan Luangthongkum (2015). "Voice register in Mon: Acoustics and electroglottography." In: *Phonetica* 72.4, pp. 237–256.

Anastasopoulos, Antonios et al. (Dec. 2020). "Endangered Languages meet Modern NLP." In: *Proc. COLING*. Barcelona, Spain (Online): International Committee for Computational Linguistics, pp. 39–45. URL: https://aclanthology.org/2020.coling-tutorials.7.

Besacier, Laurent et al. (2014). "Automatic speech recognition for under-resourced languages: A survey." In: *Speech communication* 56, pp. 85–100.

Brunelle, Marc, Duy Duong Nguyên, and Khac Hung Nguyen (2010). "A laryngographic and laryngoscopic study of Northern Vietnamese tones." In: *Phonetica* 67.3, pp. 147–169.

Brunelle, Marc, Tạ Thành Tấn, et al. (2020). "Transphonologization of voicing in Chru: Studies in production and perception." In: *Laboratory Phonology* 11.1.

Childers, D.G. and C.K. Lee (1991). "Vocal quality factors: Analysis, synthesis and perception." In: *Journal of the Acoustical Society of America* 90.5, pp. 2394–2410.

Childers, Donald G. and Ashok K. Krishnamurthy (1984). "A critical review of electroglottography." In: *Critical reviews in biomedical engineering* 12.2, pp. 131–161.

Colton, Raymond H. and Edward G. Conture (1990). "Problems and pitfalls of electroglottography." In: *Journal of Voice* 4.1, pp. 10–24.

Conneau, Alexis et al. (2020). "Unsupervised cross-lingual representation learning for speech recognition." In: *arXiv preprint arXiv:2006.13979*.

Degottex, Gilles et al. (2014). "COVAREP—A collaborative voice analysis repository for speech technologies." In: *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp. 960–964.

Do, Thi Ngoc Diep, Alexis Michaud, and Eric Castelli (2014). "Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages." English. In: *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*. Conference dates: 14-16 May 2014. St Petersburg, pp. 153–160. ISBN: ISBN 978-5-8088-0908-6. URL: http://halshs.archives-ouvertes.fr/halshs-00980431/.

Dobrovolsky, Michael and Francis Katamba (1996). "Phonology: the function and patterning of sounds." In: *Contemporary Linguistics: An Introduction. Essex: Addison Wesley Longman Limited*.

Fabre, Philippe (1957). "Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute fréquence." In: *Bulletin de l'Académie Nationale de Médecine* 141, pp. 66–69.

— (1958). "Etude comparée des glottogrammes et des phonogrammes de la voix humaine." In: *Annuaire Oto-rhino Laryngologie* 75, pp. 767–775.

— (1959). "La glottographie électrique en haute fréquence: Particularités de l'appareillage." In: *Comptes rendus des séances de la Société de biologie et de ses filiales* 153.8-9, pp. 1361–1364.

— (1961). "Glottographie respiratoire, appareillage et premiers résultats." In: *Comptes rendus hebdomadaires des séances* 252.9, p. 1386.

Gao, Jiayin (2016). "Sociolinguistic motivations in sound change: On-going loss of low tone breathy voice in Shanghai Chinese." In: *Papers in Historical Phonology* 1, pp. 166–186.

Garellek, Marc et al. (2020). "Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls." English. In: *Journal of Speech Science* 9.1. ISSN: 2236-9740. URL: https://halshs.archives-ouvertes.fr/halshs-02894375.

Gendrot, Cédric, Emmanuel Ferragne, and Anaïs Chanclu (2022). "Analyse phonétique de la variation inter-locuteurs au moyen de réseaux de neurones convolutifs: voyelles seules et séquences courtes de parole." In: *Journées d'étude de la parole 2022 (JEP 2022)*.

Gordon, Matthew and Peter Ladefoged (2001). "Phonation types: a cross-linguistic overview." In: *Journal of Phonetics* 29, pp. 383–406.

Guillaume, Séverine et al. (2022). "Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)." In: *Proc. ComputEL*. Dublin, Ireland: Association for Computational Linguistics, pp. 170–178. URL: https://aclanthology.org/2022.computel-1.21.

Hampala, Vít et al. (2016). "Relationship between the electroglottographic signal and vocal fold contact area." In: *Journal of Voice* 30.2, pp. 161–171.

Henrich, Nathalie et al. (2004). "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation." In: *Journal of the Acoustical Society of America* 115.3, pp. 1321–1332.

Herbst, Christian T. (2020). "Electroglottography –An Update." In: *Journal of Voice* 34.4, pp. 503–526. ISSN: 0892-1997. DOI: https://doi.org/10.1016/j.jvoice.2018.12.014. URL: https://www.sciencedirect.com/science/article/pii/S0892199718304612.

Herbst, Christian T., W. Tecumseh S. Fitch, and Jan G. Švec (2010). "Electroglottographic wavegrams: a technique for visualizing vocal fold dynamics noninvasively." In: *The Journal of the Acoustical Society of America* 128.5. Publisher: Acoustical Society of America, pp. 3070–3078.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization." In: *CoRR* abs/1412.6980.

Kirby, James (2020). *Praatdet: Praat-based tools for EGG analysis*. Version 0.3. URL: https://github.com/kirbyj/praatdet.

Kirby, James, Pittayawat Pittayaporn, and Marc Brunelle (2022). "Transphonologization of onset voicing: revisiting Northern and Eastern Kmhmu'." In: *Phonetica* 79.6, pp. 591–629.

Kobrock, Kristina et al. (2023). "Assessing the replication landscape in experimental linguistics." In: *Glossa Psycholinguistics* 2.1.

Kochanski, Greg P. and Chilin Shih (2003). "A Quasi-glottogram signal for voicing and power estimation." In: *Journal of the Acoustical Society of America* 114.4, pp. 2206–2216.

Kuang, Jianjing and Patricia Keating (2014). "Vocal fold vibratory patterns in tense versus lax phonation contrasts." In: *The Journal of the Acoustical Society of America* 136.5, pp. 2784–2797.

Laver, John (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Liu, Zoey, Justin Spence, and Emily Prud'hommeaux (2022). "Enhancing documentation of Hupa with automatic speech recognition." In: *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Macaire, Cécile et al. (2022). "Automatic Speech Recognition and query by example for Creole languages documentation." In: *Findings of the Association for Computational Linguistics: ACL 2022*.

Mazaudon, Martine and Alexis Michaud (2008). "Tonal contrasts and initial consonants: a case study of Tamang, a 'missing link' in tonogenesis." In: *Phonetica* 65.4, pp. 231–256.

Michaud, Alexis (2004a). "A Measurement from Electroglottography: DECPA, and its Application in Prosody." In: *Speech Prosody 2004*. Ed. by Bernard Bel and Isabelle Marlien. Nara, Japan, pp. 633–636.

— (2004b). "Final consonants and glottalization: new perspectives from Hanoi Vietnamese." In: *Phonetica* 61.2-3, pp. 119–146.

Michaud, Alexis, Oliver Adams, Trevor Cohn, et al. (2018). "Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit." In: *Language Documentation & Conservation* 12, pp. 393–429.

Michaud, Alexis, Oliver Adams, Christopher Cox, et al. (2020). "La transcription du linguiste au miroir de l'intelligence artificielle : réflexions à partir de la transcription phonémique automatique." In: *Bulletin de la Société de Linguistique de Paris* 115.1, pp. 141–166. URL: https://shs.hal.science/halshs-02881731.

Michaud, Alexis, Tuân Vu-Ngoc, et al. (2006). "Nasal release, nasal finals and tonal contrasts in Hanoi Vietnamese: an aerodynamic experiment." In: *Mon-Khmer Studies* 36, pp. 121–137.

Nguyen, Minh-Chau (2021). "Glottalization, tonal contrasts and intonation : an experimental study of the Kim Thuong dialect of Muong." en. PhD thesis. Université de la Sorbonne nouvelle - Paris III. URL: https://tel.archives-ouvertes.fr/tel-03652510 (visited on 06/02/2022).

Nguyễn, Minh-Châu, Maximin Coavoux, and Solange Rossato (Nov. 2022). "Apprentissage profond pour l'estimation du quotient ouvert à partir du signal électroglottographique." In: *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*. Ed. by Leonor Becerra et al. Marseille, France: CNRS, pp. 29–38. URL: https://hal.archives-ouvertes.fr/hal-03846833.

Orlikoff, Robert F. (1998). "Scrambled EGG: the uses and abuses of electroglottography." In: *Phonoscope* 1.1, pp. 37–53.

Partanen, Niko, Mika Hämäläinen, and Tiina Klooster (2020). *Speech Recognition for Endangered and Extinct Samoyedic languages*. DOI: 10.48550/ARXIV.2012.05331. URL: https://arxiv.org/abs/2012.05331.

Prud'hommeaux, Emily et al. (2021). "Automatic speech recognition for supporting endangered language documentation." In: *Language documentation and conservation* 15.

Ravanelli, Mirco et al. (2021). *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv: 2106.04624 [eess.AS].

Recasens, Daniel and Meritxell Mira (2013). "Voicing assimilation in Catalan three-consonant clusters." In: *Journal of Phonetics* 41.3-4, pp. 264–280.

Rodríguez, Lorena Martín and Christopher Cox (Mar. 2023). "Speech-to-text recognition for multilingual spoken data in language documentation." In: *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Remote: Association for Computational Linguistics, pp. 117–123. URL: https://aclanthology.org/2023.computel-1.17.

Thieberger, Nick (2017). "LD&C possibilities for the next decade." In: *Language Documentation and Conservation* 11, pp. 1–4. URL: http://hdl.handle.net/10125/24722.

Vaissière, Jacqueline (2011). "On the acoustic and perceptual characterization of reference vowels in a cross-language perspective." In: *The 17th International Congress of Phonetic Sciences (ICPhS XVII)*, pp. 52–59.

Vaissière, Jacqueline et al. (2010). "Multisensor platform for speech physiology research in a phonetics laboratory." In: *Journal of the Phonetic Society of Japan* 14.2, pp. 65–77. URL: https://www.jstage.jst.go.jp/article/onseikenkyu/14/2/14_KJ00007408569/_pdf.

Wisniewski, Guillaume, Alexis Michaud, and Séverine Guillaume (2020). "Phonemic transcription of low-resource languages: To what extent can preprocessing be automated?" en. In: European Language Resources Association (ELRA), p. 306. URL: https://shs.hal.science/hal-02513914 (visited on 04/25/2023).

