Hilla Pohjalainen

Tools for voice source analysis: Updated Aalto Aparat and a database of continuous speech with simultaneous electroglottography

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 27.11.2015

Thesis supervisor:

Prof. Paavo Alku

Thesis advisor:

M.Sc. (Tech) Manu Airaksinen

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Hilla Pohjalainen

Title: Tools for voice source analysis: Updated Aalto Aparat and a database of continuous speech with simultaneous electroglottography

Date: 27.11.2015 Language: English Number of pages: 6+47

Department of Signal Processing and Acoustics

Professorship: Speech communication technology

Supervisor: Prof. Paavo Alku

Advisor: M.Sc. (Tech) Manu Airaksinen

This thesis presents two tools for voice source analysis: updated Aalto Aparat inverse filtering programme, and a database of continuous Finnish speech and simultaneous electroglottography (EGG). A new glottal inverse filtering method, quasi closed phase glottal inverse filtering (QCP) has been implemented to Aalto Aparat, and usability of the programme has been improved. The results of the computations can now be transferred to other analysis programmes more efficiently. Also, a comprehensive manual of Aparat has been compiled.

The database of continuous speech and EGG contains 20 recitations of a Finnish text by 10 male and 10 female native Finnish speakers. The recitations were recorded with a headset condense microphone and EGG electrodes. The recording sessions were performed in an anechoic chamber, and the full database contains almost an hour of material. The data can be used e.g. when evaluating new GIF methods.

Keywords: Aalto Aparat, electroglottography, voice source analysis, glottal inverse filtering

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Hilla Pohjalainen

Työn nimi: Työvälineet äänilähteen analyysiin: päivitetty Aalto Aparat ja jatkuvan puheen sekä samanaikaisen elektroglottorafisignaalin tietokanta

Päivämäärä: 27.11.2015 Kieli: Englanti Sivumäärä: 6+47

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Puhekommunikaatiotekniikka

Työn valvoja: Prof. Paavo Alku

Työn ohjaaja: DI Manu Airaksinen

Tässä työssä esitetään kaksi työvälinettä äänilähteen mallintamiseen: päivitetty äänilähteen käänteissuodatusohjelma Aalto Aparat, sekä tietokanta jatkuvasta suomenkielisestä puheesta yhdessä elektroglottografisen (EGG) signaalin kanssa. Aalto Aparatiin lisättiin päivityksen yhteydessä yksi uusi käänteissuodatusmenetelmä, quasi closed phase inverse filtering (QCP), ja ohjelman käytettävyyttä parannettiin lisäämässä tuloksien tallennusvaihtoehtoja. Suodatustuloksia voi nyt siirtää entistä helpommin muihin analyysiohjelmiin. Lisäksi laadittiin kattava ohjekirja ohjelman käytöstä.

Jatkuvan puheen ja EGG signaalin tietokanta sisältää 20 nauhoitetta, joissa lyhyt suomenkielinen tekstinäyte on luettu ääneen. Lukijoina oli 10 mies- ja 10 naispuolista suomenkielistä puhujaa. Ääneenluvut tallennettiin pantamikrofonin ja EGG electrodien avulla. Äänitykset tehtiin kaiuttomassa huoneessa, ja kokonaisuudessaan tietokanta sisältää noin tunnin verran materiaalia, jota voidaan käyttää mm. uusien äänilähteen käänteissuodatusmenetelmien arvioimiseen.

Avainsanat: Aalto Aparat, elektroglottografia, äänilähteen mallintaminen, äänilähteen käänteissuodatus

Preface

I want to thank my supervisor Prof. Paavo Alku for welcoming me to do my Master's Thesis in the Speech communication group, and my instructor M.Sc. Manu Airaksinen for excellent guidance and valuable comments all the way through the process. Additional thanks to Lic.Sc.(Tech.) Tiina Murtola and M.Sc. Ilkka Huhtakallio for helping me with various technical problems.

Next, I would like to give collective thanks to my colleagues at Department of Signal Processing and Acoustics. Many of my thesis related problems have been solved during coffee breaks, and those conversations have had a great influence on my work. At the same time I want to thank my friends for reminding me that there is no point of having a Master's Degree is you are not able to improvise a scat solo. Special thanks are given to my dear father, who has always encouraged me to keep my ambitions high.

Last, I want to thank my best fiend and the love of my life Janne. Without his endless support I wouldn't be this far.

Otaniemi, 23.11.2015

Hilla Pohjalainen

Contents

Abbreviations

AME	Attenuated main excitation
AR	Autoregressive
CC / CCD	Complex cepstrum-based decomposition
ClQ	Closing quotient
CP	Closed phase
CPIF	Closed phase inverse filtering (algorithm)
DAP	Discrete all-pole model
DFT	Discrete Fourier transform
DIR	Direct inverse filtering (algorithm)
DQ	Duration quotient
EGG	Electroglottograph(ic)
FFT	Fast Fourier transform
GCI	Glottal closure instant
GIF	Glottal inverse filtering
HRF	Harmonic richness factor
IAIF	Iterative adaptive inverse filtering (algorithm)
LF	Liljencrants-Fant model
LP	Linear prediction
LPC	Linear predictive coding
MVDR	Minimum variance distortionless response model
NAQ	Normalised amplitude quotient
OQ	Open quotient
PQ	Position quotient
PSIAIF	Pitch synchronous iterative adaptive inverse filtering (algorithm)
PSP	Parabolic spectral parameter
QCP	Quasi closed phase inverse filtering (algorithm)
RQ	Ramp quotient
SQ	Speed quotient
STE	Short-time energy
TKK	Teknillinen korkeakoulu (former Helsinki University of Technology)
SWLP	Stabilised weighted linear prediction
WDAP	Weighted discrete all-pole model
WLP	Weighted linear prediction
ZZT	Zeros of the z-transform

1 Introduction

Speech is the principal mode of communication for human beings, and the modelling the speech production mechanism is an interesting yet challenging interdisciplinary topic. For computational analysis of speech, one of the key algorithms is glottal inverse filtering (GIF). GIF methods are used for estimating the voice source signal from recorded speech, and they are an essential part of fundamental research of speech and speech technology, and they are utilised also in speech synthesis, phonetics, and phoniatrics.

The estimation of the voice source signal provides information about the speech production and allows detailed analysis and modelling of the speech signal [1], and the results can be used in various different applications. However, especially when analysing large amount of data GIF can be time consuming and complicated, as the computation involves several different parameters that can all be fine-tuned separately.

The goal of this thesis is to update a glottal inverse filtering programme Aalto Aparat and present a new database of continuous speech and simultaneous electroglottographic signal. These two parts are developed to meet the need of a user-friendly tool for glottal inverse filtering and uniform evaluation material of the existing methods. They are both to be published as open source tools. Due to slightly different nature of these two matters the experimental section of this thesis consists of two parts: first, the updating process of Aalto Aparat inverse filtering programme is presented. The second part is a report of the collection of the database of continuous Finnish speech with simultaneous electroglottograph.

The thesis is structed as follows: Chapter 2 present basic information about speech production and analysis. The concept of linear prediction is explained in brief along with several other speech feature extracting methods. Chapter 3 continues with more detailed introduction to inverse filtering methods, and it is followed by a brief introduction to glottal source parametrisation in Chapter 4. Chapter 5 is dedicated to the updating process of Aalto Aparat, and the collection of a database of continuous speech is presented in Chapter 6. The final conclusions of the thesis are presented in Chapter 7. Additionally, the manual of Aalto Aparat can be viewed as an
Appendix A.

2 Voice source analysis

In speech analysis, it is essential to understand the physical and mathematical models of the human vocal system. The basics of speech production as well as its mathematical modelling techniques are presented in this chapter.

2.1 Speech production

Speech is the principal mode of communication for human beings [2]. Speech sounds produced by human vocal organs consists of phones, which combine to form syllables and again words of different languages. These combinations serve as a symbolic representation of information.

An illustration of the *human vocal organs* is shown in Fig. 1. In speaking, the air flows from the lungs through the *larynx* in to the *vocal tract*, which consists of the *pharynx* and the *oral* and *nasal cavities*. Speech sounds are produced by interrupting and modifying this flow of air with the *vocal folds* and the vocal tract before its radiation from the mouth (and nose) to the surroundings.

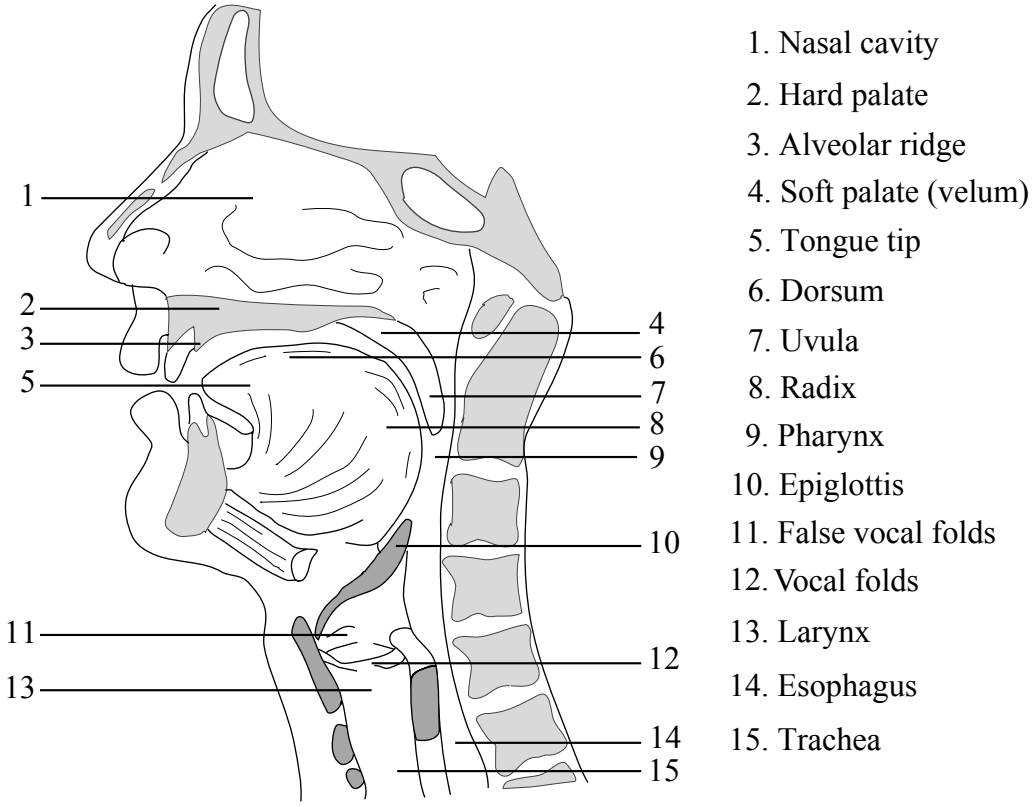


Figure 1: The human voice organs [3]

The vocal folds serve as a principal, yet not only, source of speech. The vocal folds modulate the air flow by changing the size of the *glottis*, a V-shaped gap between them. As the most typical example, when producing *voiced sounds* vocal folds vibrate

rapidly and generate quasi-periodic pulses of air that serve as a *glottal excitation* for the vocal tract (Fig. 2). In the normal (vibration) mode the vocal folds open and close completely during the cycle, but either of the phases may also dominate the cycle (e.g. incomplete closing phase produces a breathy voice, as the air flow does not stop at any point).

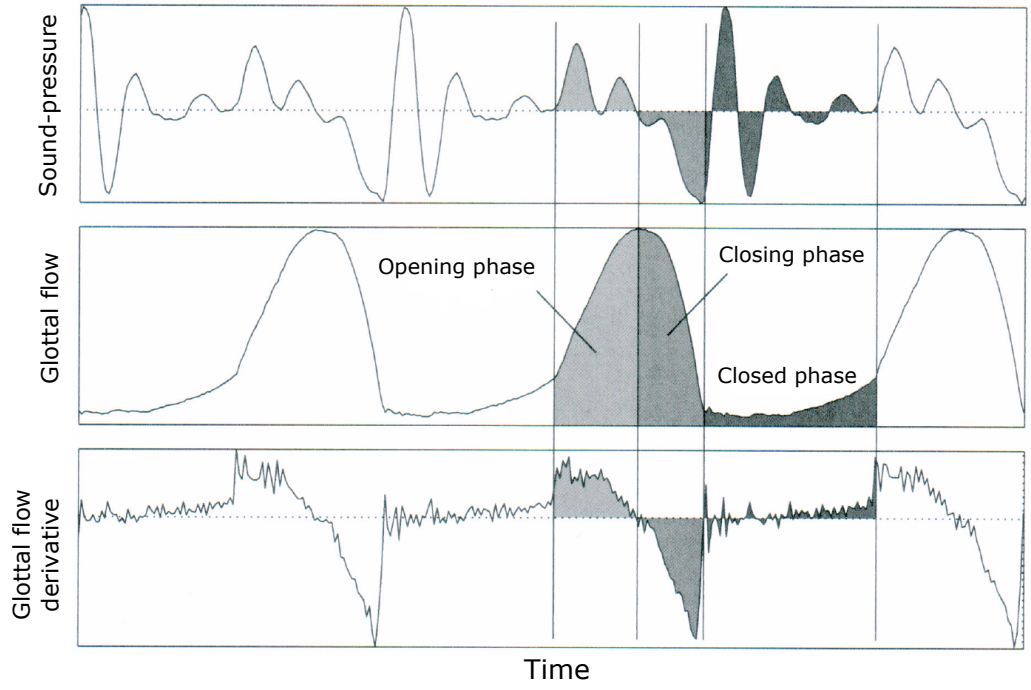


Figure 2: Three periods of a sound-pressure waveform, and the respective glottal flow and its derivative during voiced sound. The opening, closing, and closed phases of the glottal flow are highlighted. [4]

In the vocal tract the glottal excitation is transformed (filtered) into subtle sounds of speech. The vocal tract can be considered as a resonance tube for the excitation sound, and the frequency selectivity of this tube defines the spectrum of a prospective sound. The resonance frequencies of the vocal tract are called *formant frequencies* or *formants*. Varying the shape of the vocal tract by movements of the tongue, lips, and the soft palate (etc.) moves the positions of the formants and therefore produces different sounds.

Finally, the sound is radiated through the mouth (and nose). The size and the shape of the mouth opening do not have a significant effect on the transfer, but the radiation efficiency improves at higher frequencies as the wave length of the produced sound approaches the size of the opening. This so-called *lip-radiation effect* can be approximated as a rise of 6 dB per octave [2].

2.2 Parametrized spectral models of speech

Speech production can be divided into three separate processes: the glottal excitation, the vocal tract filtering and the lip radiation effect [5]. Various mathematical (and also physical) methods have been suggested for modelling this system. These methods can be used to model the speech production in general, but also for parametrisation of the vocal tract, as in glottal inverse filtering algorithms. The methods that are presented in this thesis form a filter model, that determines how the vocal tract affects the source signal. To realise this, the modelling methods try to capture the *envelope structure* of the speech spectrum. This kind of spectral estimation in an essential aspect of an automatic speech recognition and many other speech-processing algorithms [6].

An example of a spectral envelope is viewed on the Fig. 3. The envelope represents an estimation of the formants of the vocal tract, but it does not necessarily model the spectral valleys or individual peaks in a group. The objective of the estimation is to achieve sufficient accuracy that presents all of the formants but leaves the harmonic structure of the spectrum out of the envelope.

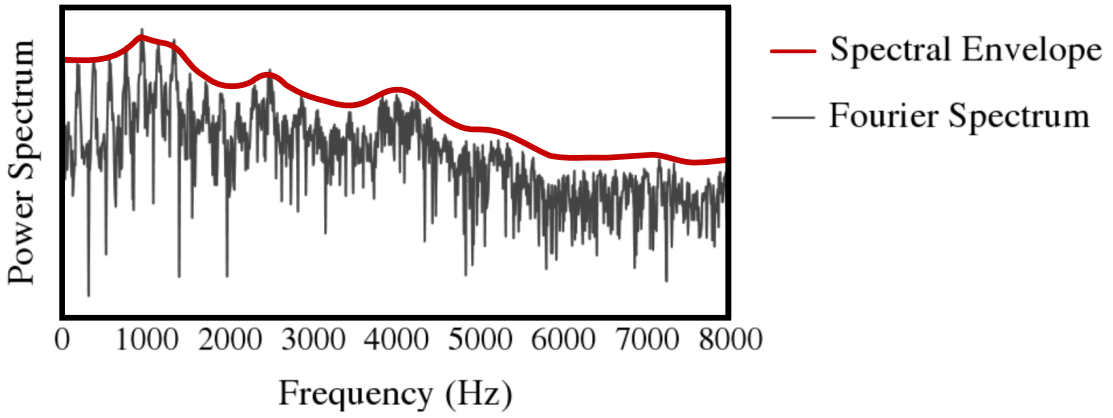


Figure 3: The spectral envelope structure and the original Fourier spectrum [6]. Peaks of the Fourier spectrum represent formants (resonance frequencies) of the vocal tract.

Many different speech feature extraction methods have been proposed over the years [6]. Three parametric methods are discussed in further detail in the following Sections 2.2.1-2.2.3., by moving on from the most popular one to the more advanced methods.

2.2.1 Linear prediction

Linear prediction (LP) analysis is a widely used method to determine the filter parameters, such as formants and spectra, of speech production [6]. LP relies on the

assumption that the original output signal s_n is *predictable* from *linear* combinations of past outputs [7]. This can be presented as an all-pole model in the Equation 1

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n, \tag{1}$$

where coefficients a_k are the linear prediction coefficients, p is the order of the prediction, and u_n is the excitation signal with G as its gain parameter. The order p of the prediction determines the level of particularity of the spectral envelope: a low order filter does not necessarily detect all the formants, but a high order filter captures also the harmonic structure of the spectrum [6]. In speech processing the recommended order is usually dependent on the sample rate F_s, being $p = F_s/1000+2$ [6].
The synthesis filter structure, i.e., transfer function $H(z)$ has only poles and is of a form

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \tag{2}$$

where $S(z)$ is the output speech and $E(z)$ the impulse excitation signal. This is also the filter function of the spectral envelope. The gain parameter G and the filter coefficients a_k must still be determined.

The coefficients a_k are solved by using the method of least squares. First, the total squared error e_n (also known as residual) is minimized. The residual signal e_n is a difference between the original and predicted signal. The original signal is assumed to be deterministic. When the total squared error is denoted as E,

$$E = \sum_n {e_n}^2 = \sum_n (s_n + \sum_{k=1}^{p} a_k s_{n-k})^2. \tag{3}$$

E is minimized by setting

$$E = \frac{\partial E}{\partial a_i} = 0, 1 \leq i \leq p. \tag{4}$$

From the equations (3) and (4) we get the *normal equations*

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = -\sum_n s_n s_{n-i}, 1 \leq i \leq p. \tag{5}$$

For any definition of the signal s_n, a set of p equations forms in p unknowns, which can be solved for the predictor coefficients a_k $(1 \leq k \leq p)$. When it is specified that the error is minimized over the infinite duration, $-\infty < n < \infty$, the normal equations take a form

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), 1 \leq i \leq p, where\ R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i}. \tag{6}$$

This method is called autocorrelation method since the coefficients $R(i-k)$ form an autocorrelation matrix. Hence, the coefficients a_k can be solved from the Equation (6) by using the *Levinson-Durbin recursion* [7].

Once the filter function has been defined, it is possible to compute the ideal excitation (*residual*) estimation signal $E(z)$. Derived from (2), the equation for the residual is

$$E(z) = \frac{S(z)}{H(z)}, \tag{7}$$

where $1/H(z)$ denotes an inverse filter. This is a common approach in e.g. glottal inverse filtering, where the vocal tract is first modelled and the model is used to determine the LP-based glottal flow estimate.

The advantages of LP are its simplicity of computation, ability to provide a spectral envelope for both voiced and unvoiced speech, and its compatibility with efficient vector quantisation techniques. However, the LP spectrum does not model high-pitched voiced speech exactly [8]. Especially bias in the formants may occur.

The LP method is the basis of many more developed modelling methods, such weighted linear prediction (WLP) [9] and stabilised weighted linear prediction (SWLP) [10]. As a difference to the LP method, the WLP method applies temporal weighting of the square of the residual signal e_n when computing all-pole models. Several different weighting functions have been developed for this, e.g. short-time energy (STE) function [9] or attenuated main excitation (AME) function [11]. The WLP method relies on the idea of emphasising those samples that fit well the speech production model, but unfortunately it does not guarantee stability of the model. As a more advanced method the SWLP utilises a modified weighting function, and provides a stable all-pole filter.

2.2.2 Discrete all-pole model

The discrete all-pole (DAP) method was developed especially for the modelling tasks where only a discrete set of spectral points is given [8]. The DAP approximates the spectrum of speech by line spectrum, and the spectral envelopes are then representable by a relatively small discrete set of values. This is achieved by utilising the Itakura-Saito error measure [12, 13]: the DAP computes the parameters of an autoregressive model by minimizing a discrete version of the Itakura–Saito distance, instead of the time squared error used by the LP. The use of Itakura-Saito distance is justified, because it measures the spectral flatness as the geometric mean of the spectral samples divided by their artihmetic mean. The optimal model makes the residual (error) spectrum as flat as possible [8].

The DAP algorithm is computationally more intensive than LP, as it iterates the coefficients a_k separately, and each iteration requires two real discrete Fourier transforms of size N (number of the discrete frequencies) and the solution of a set of $p+1$ linear equations [8]. The autocorrelation matrix R stays constant throughout the iterations, though, and therefore needs be inverted only once. As in LP model,

the matrix is Toeplitz symmetric, which allows the use of efficient algorithms in solving the coefficients.

Generally DAP modelling produces better fitting spectral envelopes than the LP model, as they are less biased towards the pitch harmonics. Additionally, the DAP modeling can also be modified to allow weighting of the error measure e_n as a function of frequency [8]. When tested with synthetic vowels, the weighted DAP, or WDAP, modeling has proven to improve the estimates for formants that are lower than the cut-off frequency of the weighting function. However, this happens at the expense of the higher formants [8]. At the same time, there are many applications of speech processing that actually desire better spectral accuracy at the low frequencies.

2.2.3 Minimum variance distortionless response model

The minimum variance distortionless response (MVDR) model is another spectrum estimation method developed to overcome the shortcomings of LP models [14]. When compared to LP, the MVDR method provides better models for medium and high pitch voiced speech: with the proper choice of filter order, the MVDR spectrum models the formants and spectral powers of voiced speech exactly.

Similarly to the LP method, the MVDR filter is also an all-pole filter that is obtained by the Levinson-Durbin computation [14]. However, in the MVDR method the filter is designed so that its response at a particular frequency w_1 has unity gain. This is achieved by limiting the output variance to the distortionless constraint. The power at the output of the optimised constrained filter is used as an estimate of the power spectrum at the frequency w_1. When this estimation is performed individually for each frequencies, the result is an exact model through the spectrum, also on the harmonic frequencies of voiced speech.

The actual computation of the power spetrum is not much more complex in MVDR model than in LP method, even if there is a need to design a separate filter for each frequency, as MVDR method is mathematically based upon the LP coefficients [14]. Moreover, the quantising methods for LP coefficients can by utilised on the MVDR method as well.

The MVDR spectrum improves as the model order increases. This can be seen as a good modelling behaviour, and it is a feature that the LP model does not have. However, the LP model can still be more convenient method on the lower pitches: the exact crossover point on medium frequencies between the two methods is still to be found [14].

3 Glottal source estimation

The functionality of human speech production mechanism is an interesting yet a difficult research subject for traditional research methods: e.g. the oscillation of the vocal folds can not be measured directly due to their hidden position, nor observed without special instruments [15]. These instruments are needed when analysing the vocal tract visually, and their usage is always invasive and might therefore affect the speech production itself. Additionally, these methods are not easily achievable as the instruments are often expensive and meant for professional usage. Hence, it is not surprising that alternative, mathematical models of the behaviour of the speech production system have been developed as well.

The glottal inverse filtering (GIF) is a computational analysis method that enables to study the voice production from an acoustical signal [15]. The idea of GIF is that the voice production is modelled as an *inverse problem*, where the known speech signal output is used to estimate the unknown input signal, the glottal excitation. This is achieved by modelling the filtering effects of the vocal tract and lip radiation, and then cancelling these effects from the speech. In other words, the speech is filtered through the *inverses* of the vocal tract and lip radiation models.

Glottal flow estimation is one of the basic problems of speech processing [16], and six techniques to solve it are reviewed in this chapter. Methods presented in subsections 3.1-3.3 are based on inverse filtering, while the methods in 3.4 rely on the mixed-phase properties of speech.

3.1 Closed phase inverse filtering

The methods based on closed phase inverse filtering (CPIF) compute the estimate of the vocal tract during the closed phase of the glottal vibration [17], typically by utilising LP analysis with the covariance criterion over speech samples of the closed phase. The advantage of the method is that the effects of the subglottal cavities are minimized from the vocal tract transfer function. However, an accurate determination of the closed phase is not simple, especially on a high-pitch voices where both open and closed phases are so short that they do not contain enough samples for the estimation. Several techniques are proposed to solve this: e.g. multi-cycle closed-phase LPC [18], where a filter estimation is compiled of samples from a small amount of neighbouring glottal cycles. Another approach is to allow non-zero glottal wave to exist over closed glottal phases [19].

The analysis procedure of CPIF is visualised in Fig. 4. First, the digitized speech samples are passed through a linear-phase high-pass filter to remove any low-frequency energy. Second, sequential covariance method analysis is performed and normalized squared error is computed for each frame. The closed phase is determined by these error values, and only the closed time span is taken into account when creating the (traditionally LP-based) vocal tract model. Finally, the glottal flow derivative estimate is obtained by inverse filtering the input speech signal with the inverse of the vocal tract model.

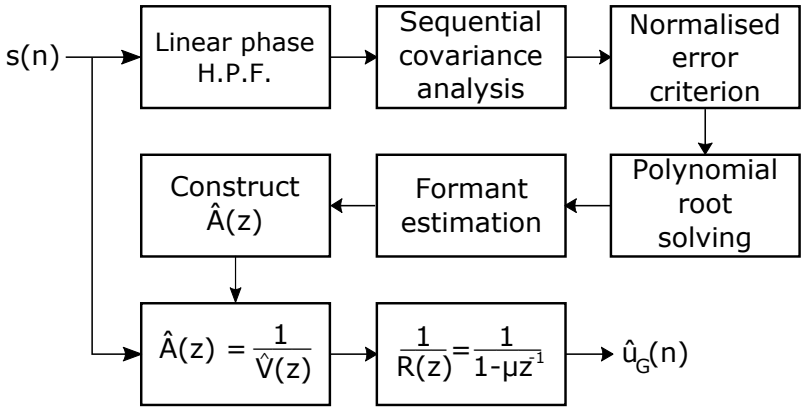


Figure 4: The block diagram of CP method. Â stands for inverse vocal tract model estimate. [17]

3.2 Quasi closed phase glottal inverse filtering

The quasi closed phase (QCP) inverse filtering is based on the principles of CP analysis, but in contrast with the basic CP method, the estimation are applied with the autocorrelation criterion using a long analysis window [20]. This is enabled by weighted linear prediction (WLP) [9] with an attenuated main excitation (AME) weighting function [11].

The algorithm consists of four stages (Fig. 5). First, glottal closure instants (GCIs) [21] are detected from the speech signal to determine the closed phase. This can be performed by using electroglottography or specific GCI detection algorithms. After that, the AME weighting function is constructed based on the estimated GCIs. The AME function contains three parameters: the position quotient (PQ), the duration quotient (DQ), and the ramp quotient (RQ, number of transitional samples of the AME window) (Fig. 6). The AME function is utilised when creating the WLP based vocal tract model.

3.3 Iterative adaptive inverse filtering

The iterative adaptive inverse filtering (IAIF) method [23] is based on an iterative refinement of both the vocal tract and the glottal components. In IAIF the glottal flow and its contribution to the speech spectrum is estimated with an iterative structure, which is repeated twice during the computation (Fig. 7). The final result (as well as the first estimate) is obtained by cancelling the effects of the vocal tract and lip radiation from the original speech signal by inverse filtering.

IAIF method presented in most scientific articles utilizes linear predictive analysis (LPC) to estimate the vocal tract. However, e.g. DAP [8] and MVDR models [6] can be used as well.

IAIF is closely connected with the pitch synchronous iterative adaptive inverse filtering (PSIAIF) [23] and the direct inverse filtering (DIR) [4] methods. In PSIAIF, the IAIF analysis is applied twice for the same signal. The first application gives

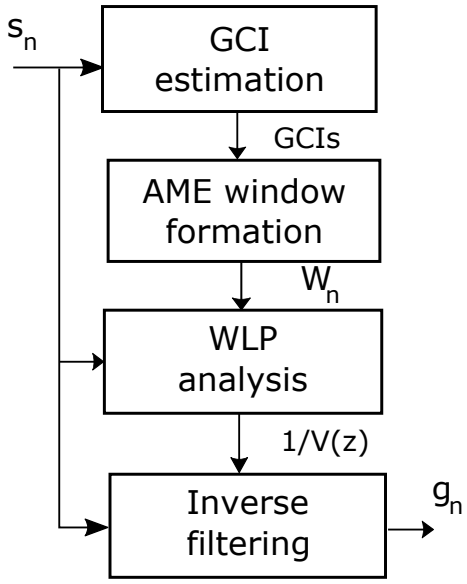


Figure 5: The block diagram of QCP method [20]

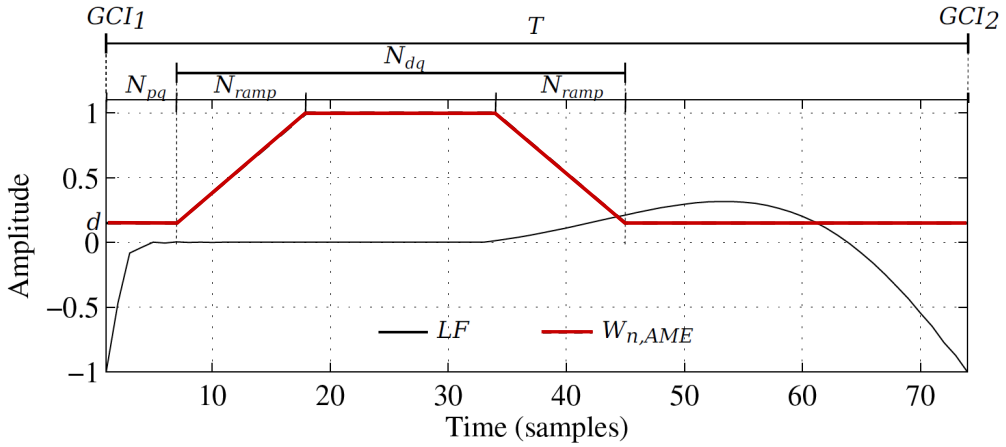


Figure 6: A glottal flow derivative waveform (LF) and the AME weight function $W_{n,AME}$ between two consecutive GCIs. Glottal flow derivative was computed from the LF model [22]. The AME function was computed with PQ = 0.1, DQ = 0.5, and RQ = 10 [20]

the glottal wave estimate that spans over several pitch periods, while the second application is over one glottal cycle only (from maximum glottal opening to the next maximum). This has proven to improve the final estimate, as the speech spectrum can be kept free from the harmonic structure of the source [23].

Instead, the DIF method is actually a simple version of IAIF: the algorithms of the two methods are identical until the first iteration, after which the DIR stops the process and gives the first estimation $g_1(n)$ as the final result (Fig. 7) [4]. This

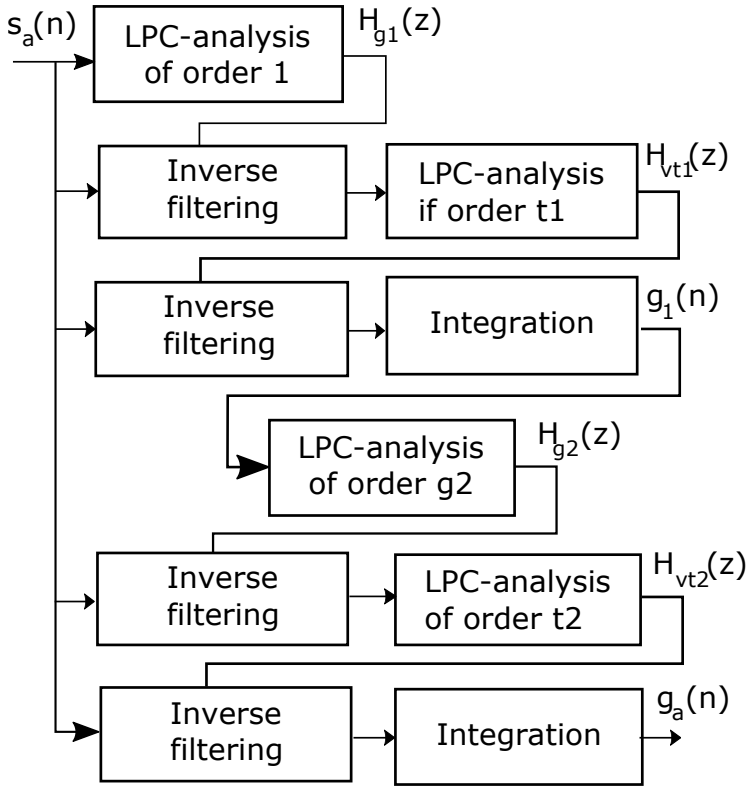


Figure 7: The block diagram of IAIF method. $H_{g1}(z)$ and $H_{g2}(z)$ represent LPC models, $H_{vt1}(z)$ and $H_{vt2}(z)$ the vocal tract filters, and $g_1(n)$ and $g_a(n)$ are the glottal flow estimates of the first and second phases, respectively [23].

makes the DIF fast, yet not as accurate method as the IAIF.

3.4 Mixed-phase models

The mixed phase models are based on the idea that the speech is a convolution of a maximum phase glottal excitation signal with a minimum phase vocal tract filter impulse response [24]. Both of the signals are assumed to be stable, but by their nature the glottal signal is assumed to be anti-causal while the vocal tract filter is assumed to be causal. The resonances of the two signals can be detected separately by using differential spectra of z-transform: these anti-causal poles outside of the unit circle on z-plane correspond to resonances of the glottal source signal, while the causal poles correspond to the vocal tract resonances.

The source and tract formants are detected from the speech signal by processing several differential phase spectra calculated on various circles on the z -plane, and then grouping these spectra into two sets according to the placement of the poles (outside or inside the unit circle). Frequencies of resonances are tracked by inspecting positive and negative peaks on the differential phase spectra: each peak stand for one resonance, and its sign indicates the causality. Positive peaks are inside the unit circle and therefore classified as vocal tract resonances, while negative peaks indicate

anti-causal glottal excitation signal.

Example of such a decomposition methods are zeros of the z-transform (ZZT) [25] and complex cepstrum-based decomposition (CC / CCD) [26]. In both methods speech frames are first weighted by a specific window (e.g. Hanning or Hamming) [26]. In the ZZT domain, the unit circle is used as a discriminant boundary, while the complex cepstrum domain utilises the frequency origin to separate the minimum and maximum phase components.

4 Glottal source parametrisation

When the glottal source have been estimated by GIF analysis, the next stage is typically the parametrisation of the estimated glottal excitations: the obtained waveforms are expressed numerically with properly selected *glottal flow parameters* [15]. The purpose of these parameters is to select the most important features of the glottal source waveforms, and represent them in a compressed numerical form. Different parametrisation methods focus on different aspects of the behaviour of the glottal function, and hence the selection of the parametrisation method has a significant impact on the final results of the research problem in question. It is therefore essential to understand the functionality of different alternatives before the selection of the parametrisation method.

In this chapter different types of parametrisation methods of glottal source are briefly presented. They are divided into two subsections: time-domain and frequency-domain methods.

4.1 Time-domain parametrisation methods

The most common, classical time-based parameters to represent glottal excitation waveforms are the open quotient (OQ), the closing quotient (ClQ), and the speed quotient (SQ) [27, 28]. The names refer to the phases of glottal cycle, as all of these quotients are defined as the ratio of the time-durations of different phases (Fig. 8). For example, the closing quotient is the ratio of the closing phase to the duration of the fundamental period. The starting and ending points of each phase can be extracted either manually or automatically from the excitation waveforms.

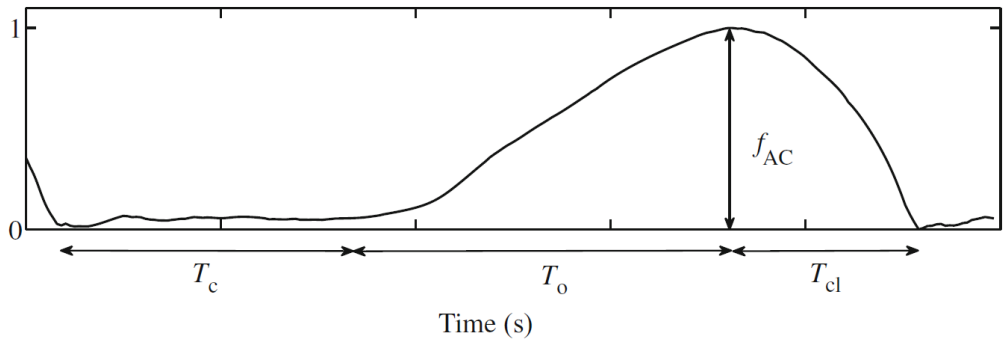


Figure 8: Computation of time-based parameters from the glottal pulse, with time-durations marked: closed phase (T_c), opening phase (T_o), and closing phase (T_{cl}). Classical time-based parameters, open quotient (OQ), speed quotient (SQ), and closing quotient (ClQ) are defined from these values as follows: OQ $= (T_o + T_{cl})/T$, SQ $= T_o/T_{cl}$, and ClQ $= T_{cl}/T$, where $T = T_c + T_o + T_{cl}$ [15].

Despite the simple concept of the classical time-based quantities, the computation of them is often problematic: formant ripple or noise in the estimated glottal excitation may occur, and detecting the exact time instants becomes challenging. Therefore, the

true time instants are sometimes replaced by other points of the cycle: e.g. instants that mark a certain ratio of the minimum to maximum amplitudes of the glottal cycle [29].

The time-based parameters can be quantified also by measuring the amplitude-domain values, such the AC-amplitude of the flow and the negative peak amplitude of the glottal flow derivative. These two values are the extreme values of the flow and its derivative, and therefore easily extracted from the glottal source waveforms even if the waveform was not perfectly noiseless. The normalised amplitude quotient (NAQ) is an example of this kind of amplitude-domain value using, yet still time-domain based quotient [30].

4.2 Frequency-domain parametrisation methods

Parametrisation of the glottal excitation can also be approached by utilising frequency domain measures, since the time-domain variation (e.g. in phonation) can be viewed on the frequency domain as an alternation in the decay of the spectral envelope of the glottal excitation. The source spectrum is typically computed using the fast Fourier transform (FFT), but also parametric spectral models, e.g. auto-regressive (AR) spectral estimation methods can be used [15].

Among all the parametrisation methods for the voice source spectrum, the most straightforward one is the alpha ratio, which is defined as a spectral ratio between spectral energies below and above a certain frequency [31]. However, the more justified method is to compute the spectral decay of the glottal source after the level of the fundamental frequency (F0) and its harmonics. Harmonic richness factor (HRF) [32] is an example quotient of this kind of approach: HRF is defined from the spectrum of the estimated glottal source as the ratio between the sum of the amplitudes of harmonics above the fundamental frequency F0 and the amplitude of the fundamental frequency itself.

HRF is not the only method utilising the spectral harmonics of the glottal source in the quantisation of voice production: the source spectrum can be analysed also by linear regression analysis over the first eight harmonics [33], or, computing the difference between the amplitudes of the fundamental and the second harmonic [34]. This later approach is usually denoted as H1-H2 (or H12).

As an alternative to the methods mentioned above, the measure of the voice spectrum can also be based on a pitch synchronously computed spectrum: the the parabolic spectral parameter (PSP) matches a second-order polynomial to the flow spectrum over a single glottal cycle [35]. The measure has been shown to be effective on differentiating phonation types of varying spectral slopes.

5 Updating Aalto Aparat

A MATLAB-based Aalto Aparat (later Aparat) is a voice inverse filtering and parametrisation software for estimating and analysing the glottal flow. In this chapter the history and the updating process of the programme are described as well as the new features of the software are presented.

5.1 The original TKK Aparat

The original TKK Aparat was created and designed by Matti Airas [4] as a part of his doctoral thesis. In his research on vocal emotions in the vowel segments of continuous speech [36], there was a need to analyse the glottal flow by glottal inverse filtering as the natural emotion detection turned out to be fairly challenging. Emotion analysis requires a large amount of data and a possibility to manipulate and fine-tune different inverse filtering parameters individually, which makes it a time consuming work without the proper tool. Therefore Aparat was created to meet this demand, and it has later proven to be a useful tool not only in emotion analysis of speech but also in algorithm development, other speech science research, and the study of occupational voice.

The TKK Aparat included three GIF methods: iterative adaptive inverse filtering (IAIF; the default method of the programme), direct inverse filtering (DIF), and experimental inverse filtering method. The DIF method is basically a simplified version of the IAIF method, as the DIF method follows the same filtering process as IAIF but finishes after the first integration step [4]. As the experimental method was never truly finished, these two represent also the only working methods of the programme. In addition to a user-friendly graphical interface (Fig. 9) the programme provided an easy way to save the GIF results for later use or other projects.

The first version of the software was published in 2005 under the name TKK Aparat, referring to the former Helsinki University of Technology (which was merged into Aalto University in 2010). It was available under an open source license on the *Sourceforge.com* website. Aparat is a MATLAB-based program and therefore easy to adapt e.g. in engineering where MATLAB is a widely used software. However, MATLAB utility programs for Windows and Linux operating systems were also published and available for those scientists who do not necessarily have the MATLAB license.

5.2 Update process in general

Before this master's thesis, Aparat had had its latest update on 2008 and then been left untouched until present. However, the Department of Signal Processing and Acoustics still received comments and inquieries about the programme. Even more importantly, a new more advanced GIF method had been developed since the first release, so it was justified to modernise the programme. The updating process and the documentation of it was included to this master's thesis work.

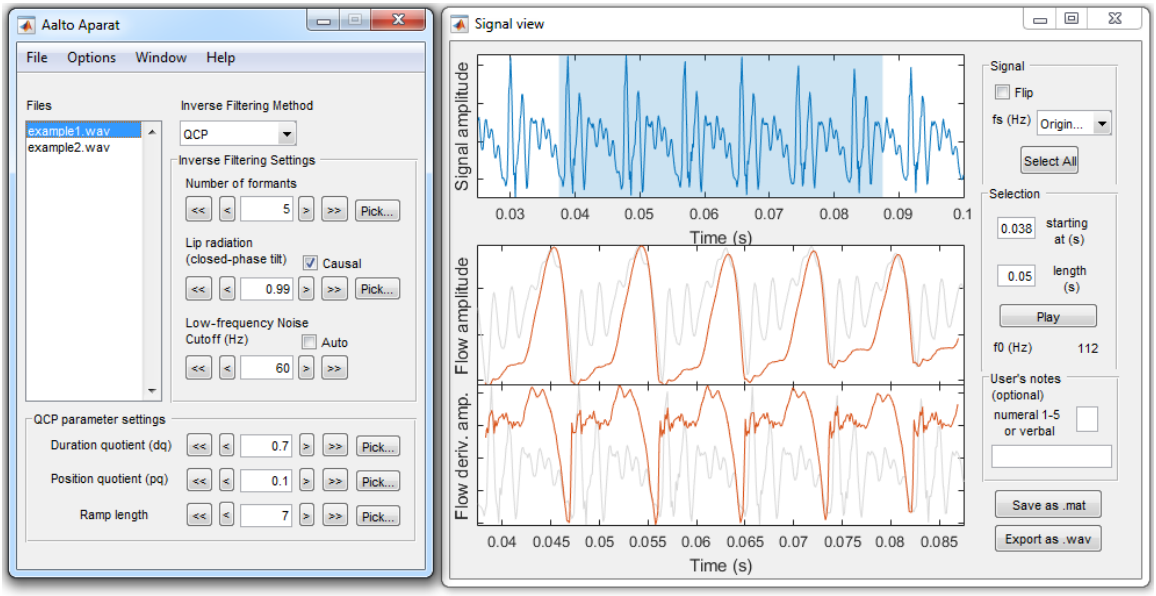


Figure 9: Aalto Aparat (former TKK Aparat) consists of two main windows: control window on the left and signal view on the right.

The version 2.0 update started with fairly mechanical debugging as some former standard MATLAB syntaxes had changed by 2015, and therefore certain functions of Aparat did not run at all. After this, the actual development and implementation of new features could finally start.

The most significant modification of the programme was the rearrangement of the GIF methods in use: two previous methods were removed, and one new method implemented and set as the current default method. As the methods do not utilise the exact same parameters, one of the main questions at this point was where to place all the different buttons and switches on the user interface. In the original version of Aparat, the problem had been solved by using an extra advanced settings panel that could be found on the drop-down menu bar of the control window. However, as this did not seem the most intuitive place, all the advanced parameter setting handles were moved to the control window right under general filtering settings (Fig. 10). This lead to noticeable changes on user interface, yet the visual style of the old interface was preserved, and the functionality of similar looking buttons and switches was kept the same.

In addition to a new GIF method, there were several minor details in the user interface that were modified according to the feedback on the original version. The functionality of signal selection and saving tools were improved to match the most common needs of the users, and some experimental yet not finished nor working tools were deleted.

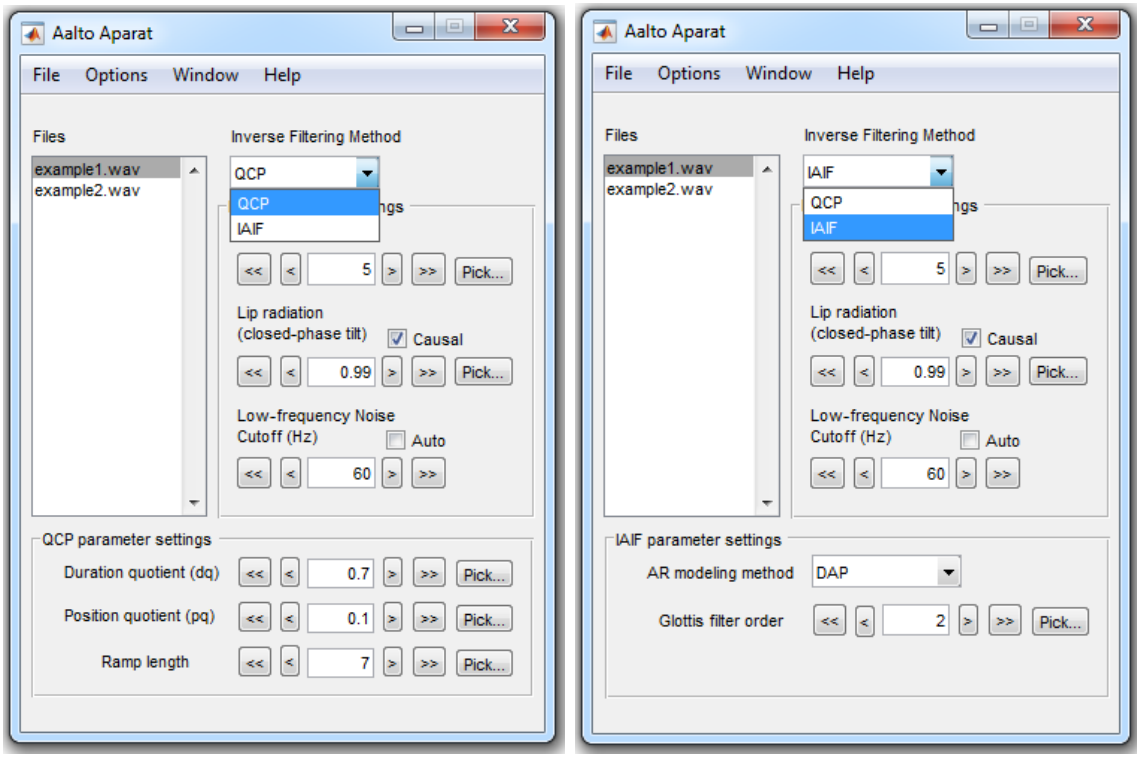


Figure 10: QCP and IAIF parameter settings appear under general parameter settings

5.3 New features of the software and the final product

The main modification to Aparat was the implementation of the QCP glottal inverse filtering method (explained in the Sec. 3.2). QCP is now the new default GIF method in Aparat, as IAIF was left as alternative method. While QCP has been shown to improve the accuracy of GIF when compared to results obtained with IAIF, it is not as robust as IAIF: especially with a large amount of data IAIF becomes more convenient tool to work with.

Another modification is the possibility to save GIF waveforms from the glottal estimate not only as .mat files but also as .wav files. This change was suggested by the users, as well as the new sample options that allow the user to select the position of the sample sequence exactly.

Despite the fact that Aparat has already been available as an open source software a full manual of its operation had never been written. The last task in the whole update process was to write the manual and attach it to the programme package. Aalto Aparat manual can be found also as an Appendix A of this thesis.

Aparat remains an open source software, but the distribution of the programme package will be moved under control of Aalto University Department of Signal Processing of and Acoustics.

6 Database of continuous speech

The other sub-project of this master's thesis was to collect a database of a continuous Finnish speech and corresponding electroglottographic signal. In this chapter the background of the project, recording practises and the final database are presented.

6.1 Electroglottography

The electroglottograph (EGG) is a device that consists of an EGG pre-amplifier and two electrodes placed external to the larynx (Fig 11). The output signal provides a measure of vocal fold contact: impedance of the substance between the electrodes varies depending on the phase of the vibration of the vocal folds [37]. When the glottis is open the impedance is high, as a high frequency current between the electrodes decreases. Respectively, when the glottis is closed, the impedance is low.

Figure 11: Electroglottograph pre-amplifier and electrodes

The electroglottogram provides complementary information to the recorded speech: the maximum of the glottal flow occurs during the open phase while the maximum of the EGG signal occurs during the closed phase. An EGG signal example can be seen on the Figure 12.

6.2 Background and former recordings

The GIF methods presented earlier in this thesis are not the only existing methods: new methods are developed and old ones improved constantly. The accuracy of the methods is tested by using for example synthetic vowels [16] or speech signals accompanied by EGG signals [38]. In the recording sessions made for this thesis

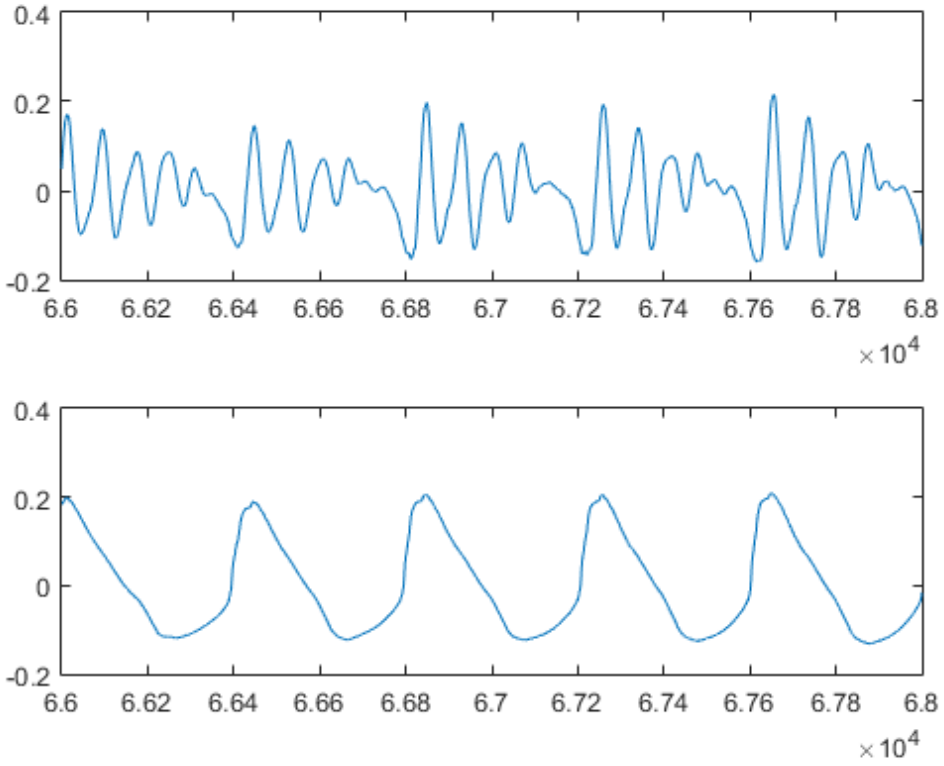


Figure 12: Example of an /a/ sound of a Finnish male. Top row presents the microphone signal, bottom row EGG signal. (Y-axis - amplitude; x-axis - time in seconds)

both continuous speech and the impedance function given by EGG were recorded at the same time. They are meant to become a part of a larger GIF method testing database in the future.

Recordings of continuous Finnish speech without EGG signal have been performed several times with a small amount of speakers, for example by Matti Airas in 2007 [39]. The same Finnish text generated and used then was selected to this new recording session as well, as in the text there are multiple [a] vowels with different levels of prominence suitable for inverse filtering [39].

6.3 Recording session practices

Speech and EGG signal of 20 native Finnish speakers were recorded. There were 10 male and 10 female participants among the subjects, and their ages ranged from 19 to 28 years, mean being 23 years. One of the speakers smoked irregularly, while the others were non-smokers. The recordings were performed in an anechoic chamber of Aalto University Department of Signal processing and acoustics.

The speakers were sitting, reciting the text from a paper attached on a folder.

They were equipped with a DPA 4065-BL headset condense microphone and electrodes for EG2-PCX2 electroglottograph [40] (Fig. 13). The microphone was placed at a 5 cm distance from the center of the lips (measured with a ruler), and the positions of the electrodes were adjusted by following the electrode placement indicator of the pre-amplifier. Both of the signals were routed through a RME Babyface sound card [41] to a laptop (Fig. 14), where they were recorded on Audacity programme with a sampling rate 48 kHz. The speech material consisted of three passages of Finnish text describing past weather conditions (Appendix B). The subjects were asked to read the text three times, with short pauses between the passages. An average recording session took slightly less than three minutes, which when multiplied by 20 participants makes about one hour recorded material.

Figure 13: EGG electrodes and microphone on a speaker

The recording set up can be seen on figure 15. The experimenter assisted in the chamber as especially EGG signal and the position of the electrodes needed continuous monitoring during the recording session. The recordings were not processed afterwards unless there were retakes or other unintended material on them. Most of the speaker managed to read the whole text without any need to a stop and start a paragraph a new.

None of the subjects had participated on a record session of this kind before, and it took relative much time to adjust the gain levels for each individual speaker, as they were all asked to speak with their natural volume and tone. In general, male participants tended to speak louder than female subjects and therefore their voice and EGG signals did not need as much amplifying as females'. It is also possible that this new situation caused a slight timidity on some of the subjects, especially as the experimenter was also in the same room listening and monitoring the performance, and they did not really dare to speak as in an everyday conditions. However, some of the subjects had experience as a radio journalist, so the reading task itself was not totally new for them.

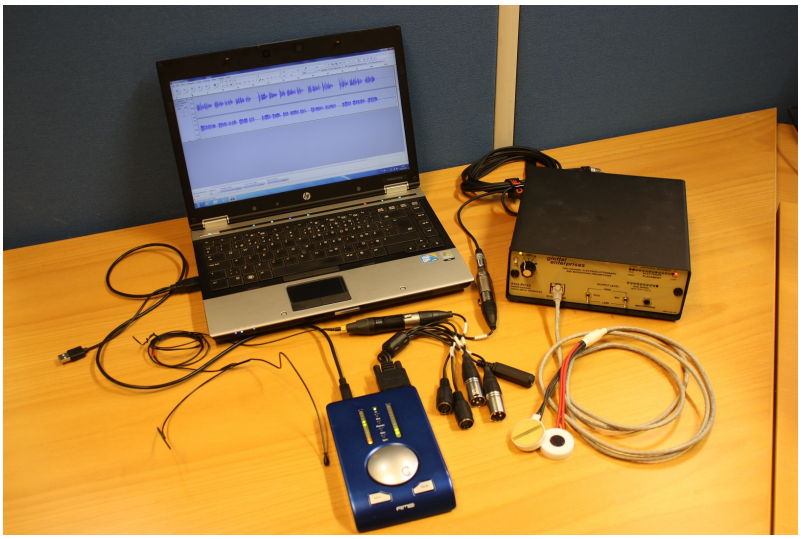


Figure 14: The equipment of the recording session: microphone, a sound card, EGG and electrodes, and Audacity programme on a laptop.

The most challenging part was to get a reasonably powerful EGG signal of every subject. Personal physiological characteristics such as the amount of adipose tissue in the neck area affect the connection as well as the volume of speech.

Figure 15: The subject (behind) is reading while the experimenter (in front) is monitoring the signals. The sound card and EGG pre-amplifier are placed on the chair next to the experimenter. Photographer: Christoffer Kauppinen

7 Summary

This thesis presented an updating process of Aalto Aparat inverse filtering programme and a recording project of continuous Finnish speech and simultaneous electroglottography. The purpose of both of the experimental projects was to provide new and increasingly better tools for voice source analysis in future.

The former TKK Aparat was updated to Aalto Aparat, as the necessary MATLAB-based and functional modifications were implemented: outdated commands were replaced and the default GIF method was changed to QCP, which has proven to improve the quality of GIF when compared to the old default, now secondary, method IAIF. Modest changes in the user interface make the programme more straightforward to use than the former version, and the new saving options enable the more diverse possibilities to store the data and utilise it in other projects as well.

Like the former version TKK Aparat, also the updated version will be distributed as an open source license. As a new part of the download package there is a newly written manual, that presents the usage and features of the programme as well as a brief theory behind the implemented methods. The manual will be part of the package from now on, and it can be updated separately in the future according to feedback from the users.

The second experimental part, the collection of a database of continuous Finnish speech and simultaneous EGG, was realised by recording sessions in an anechoic chamber of Aalto University Department of Signal Processing and Acoustic. The subjects were asked to read a short, three paragraphed text three times in a row. The recitation of 10 male and 10 female speakers were recorded with a headset microphone and EGG electrodes. The data consists of almost an hour of material, and both the audio and EGG samples have been confirmed to be of uniform quality and ready to apply.

The text read in the recordings contains multiple [a] vowels with different levels of prominence, which makes the database an useful tool for inverse filtering related research in the future. Together with EGG signals the audio data can be used when evaluating totally new, more accurate GIF methods.

References

[1] T. Raitio. *Voice source modelling tehcniques for statistical parametric speech synthesis*. PhD thesis, Aalto University, 2015.

[2] T. Rossing, R. Moore, and P. Wheeler. *The Science of Sound*, volume 3. Pearson Education, Inc, 2002.

[3] M. Karjalainen. *Kommunikaatioakustiikka*. Teknillinen korkeakoulu, Espoo, 2009.

[4] M. Airas. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008.

[5] G. Fant. *Acoustic Theory of Speech Production*. Mouton, 1960.

[6] M. Wölfel and J. McDonough. *Distant Speech Recognition*. John Wiley & Sons, Ltd, 2009.

[7] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1974.

[8] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *Signal processing, IEEE Transactions on*, 39(2):411–423, 1991.

[9] C. Ma, Y. Kamp, and L. Willems. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1):66–81, 1993.

[10] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku. Stabilised weighted linear prediction. *Speech Communication*, 51(5):401–411, 2009.

[11] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical society of America*, 134(2):1295–1313, 2013.

[12] F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electrononics and Communication in Japan*, 53-A:36–43, 1970.

[13] R. McAulay. Maximum likelihood spectral estimation and its application to narrow-band speech coding. *IEEE Transactions in Audio, Speech and Language processing*, ASSP-32(2):243–251, 1984.

[14] M. Murthi and B. Rao. Minimum variance distortionless response (MVDR) modeling of voiced speech. In *Acoustics, Speech and Signal Processing*, pages 1687–1690, 1997.

[15] P. Alku. Glottal inverse filtering analysis of human voice production - a review of estimation and parameterization methods of the glottal excitation and their applications. *Sādhanā*, 36(5):623–650, 2011.

[16] T. Drugman, B. Bozkurt, and T. Dutoit. A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26(1):20–34, 2012.

[17] D. Wong, J. Markel, and A. Gray Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *Audio, Speech, and Language Processing, IEEE Transactions on*, 27(4):350–355, 1979.

[18] D. Brookes and D. Chan. Speaker characteristics from a glottal airflow model using glottal inverse filtering. *Institute of Acoustics*, 15:501–508, 1994.

[19] H. Deng, R. Ward, M. Beddoes, and M. Hodgson. A new method for obtaining accurate estimates of vocal tract filters and glottal waves from vowel sounds. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 14:445–455, 2006.

[20] M. Airaksinen, T. Raitio, B. Story, and P. Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE Transactions in Audio, Speech and Language processing*, 22(3):596–607, 2014.

[21] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):994–1006, 2012.

[22] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):1–13, 1985.

[23] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

[24] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation using differential phase spectrums. In *ISCA ITRW VOQUAL 03*, pages 21–24, 2003.

[25] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE signal processing letters*, 12(4):344–347, 2005.

[26] T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proceedings of Interspeech*, pages 116–119, 2009.

[27] R. Timcke, H. von Leden, and P. Moore. Laryngeal vibrations: measurements of the glottic wave. *The Archives of Otolaryngology*, 68:1–19, 1958.

[28] R. Monsen and A. Engebretson. Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, 62(4):981–993, 1977.

[29] C. Dromey, E. Stathopoulos, and C. Sapienza. Glottal airflow and electroglottographic measures of vocal function at multiple intensities. *Journal of Voice*, 6(1):44–54, 1992.

[30] P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *The Journal of the Acoustical society of America*, 112(1):701–710, 2002.

[31] B. Frøkjær-Jensen and S. Prytz. Registration of voice quality. *Brüel & Kjær Technical Review*, 3(3):3–17, 1973.

[32] D. Childers and C. Lee. Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical society of America*, 90(5):2394–2410, 1991.

[33] P. Howell and M. Williams. Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech. *The Journal of the Acoustical society of America*, 91(2):1697–1706, 1992.

[34] I. Titze and J. Sundberg. Vocal intensity in speakers and singers. *The Journal of the Acoustical society of America*, 91(5):2936–2946, 1992.

[35] P. Alku, H. Strik, and E. Vilkman. Parabolic spectral parameter – a new method for quantification of the glottal flow. *Speech Communication*, 22:67–79, 1997.

[36] M. Airas and P. Alku. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, 63(1):26–46, 2006.

[37] Glottal Enterprises, Inc., 1201 E. Fayette Street. Syracuse, NY 13210. *User manual: Two-channel Electroglottograph & Microphone Amplifier. Model EG2-PCX2*, 2011.

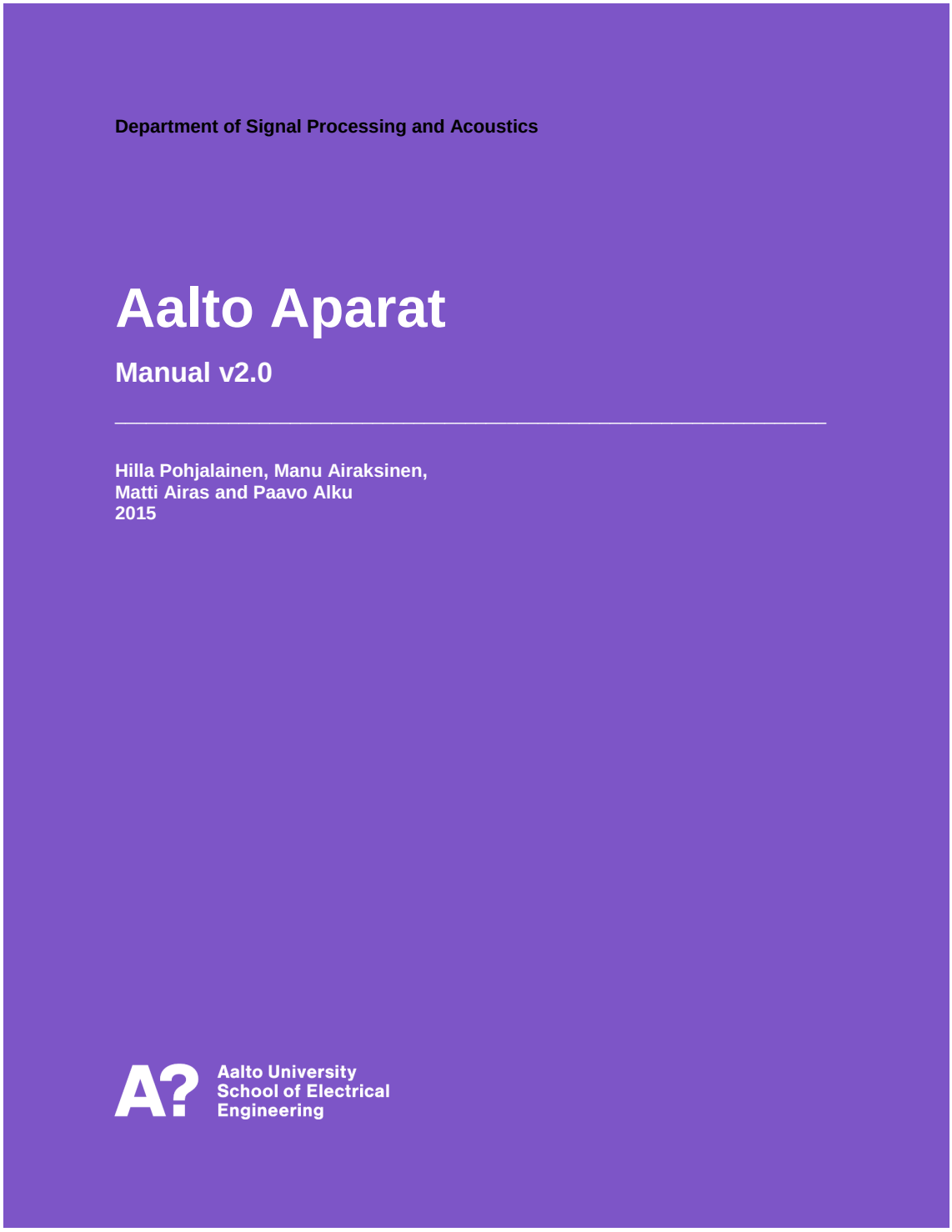
[38] A. Krishnamurthy and D. Childers. Two-channel speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:730–743, 1986.

[39] M. Airas, P. Alku, and M. Vainio. Laryngeal voice quality changes in expression of prominence in continuous speech. In *Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions in Biomedical Applications (MAVEBA)*, page 135–138, 2007.

[40] EG2-PCX2 electroglottograph home page. http://www.glottal.com/Electroglottographs.html. Accessed 11.11.2015.

[41] RME Babyface sound card home page. http://www.rme-audio.de/en/products/babyface.php. Accessed 11.11.2015.

A Aalto Aparat Manual



Contents

1 About Aalto Aparat

Aalto Aparat (later Aparat) is a voice inverse filtering and parametrisation software for estimating and analysing the glottal flow. MATLAB-based Aparat is available under an open-source licence and can be used in fundamental research of speech, fonetics and study of occupational voice. The programme provides a user-friendly graphical interface from where the results can be easily saved and re-used in other projects.

The original Aparat was created and designed by Matti Airas [1]. The first version of the software was published in 2005 under the name TKK Aparat, referring to the former Helsinki University of Technology (which was merged into Aalto University in 2010). The updated version 2.0 includes a new glottal inverse filtering (GIF) method and several minor new features, such as more versatile saving options.

1.1 Glottal inverse filtering methods

There are two different GIF methods available in Aparat: quasi closed phase analysis (QCP) [2] glottal inverse filtering and iterative adaptive inverse filtering (IAIF) [3]. The methods share three general parameters: the number of formants, lip radiation coefficient, and low-frequency noise cutoff (in Hz).

QCP is the default GIF method in Aparat since the version 2.0 update. QCP is based on the principles of closed phase (CP) analysis [4], that is the estimation of the vocal tract during the glottal closed phase. In contrast with the basic CP method, the estimation is applied with the autocorrelation criterion using a long analysis window. This is enabled by weighted linear prediction (WLP) [5] with an attenuated main excitation (AME) weight function [6].

The algorithm consists of four stages. First, glottal closure instants (GCI) are detected from the speech signal by using electroglottography or specific GCI detection algorithms. To perform this, Aparat uses an automatic algorithm called SEDREAMS [7]. After that, the AME function is constructed based on the estimated GCIs (Fig. 1). From the Aparat users' perspective, this is the most interactive part of the computation as there are three AME function parameters that the user can manipulate on Aparat control panel: position quotient (PQ), duration quotient (DQ), and the number of linear ramp samples (Nramp). The AME function is utilised when creating the WLP-based vocal tract model, and the default Aparat parameter values of PQ, DQ and Nramp have been selected in order to minimise the estimation error on a large data set of synthetic sustained vowels. In most cases, the modification of these parameters can be viewed as fine tuning, and commendable results can be obtained even without modifying them at all.
Finally, the glottal flow derivative estimate is obtained by inverse filtering the input speech signal with the vocal tract model.

IAIF is familiar to Aparat users since the first version of the software, and it was the default GIF method until QCP was implemented. While QCP has been shown

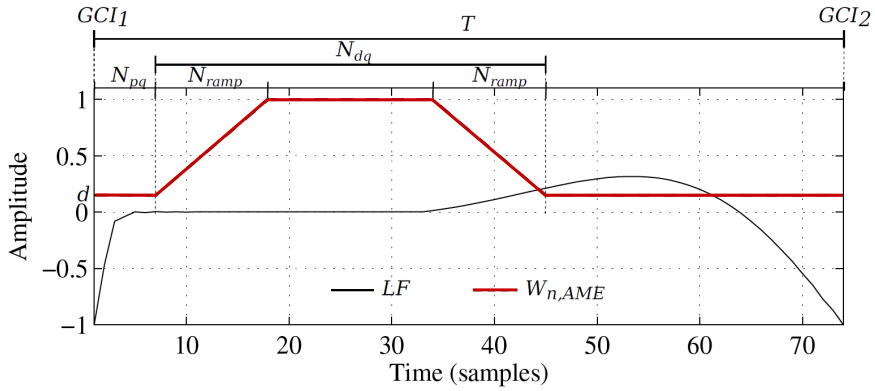


Figure 1: A glottal flow derivative waveform (LF) and the AME weight function ($W_{n,AME}$) between two consecutive GCIs. Glottal flow derivative was computed from the LF model [8]. The AME function was computed with PQ = 0.1, DQ = 0.5, and RQ = 10 [2].

to improve the quality of GIF when compared to results obtained with IAIF, it is not as robust as IAIF when reliable GCI estimates are unavailable: for example for signals with low SNR.

In IAIF method the glottal flow and its contribution to the speech spectrum is estimated with an iterative structure, which is repeated twice during the computation. The final result (as well as the first estimate) is obtained by cancelling the effects of the vocal tract and lip radiation from the original speech signal by inverse filtering.

The IAIF method presented in most scientific articles utilizes linear predictive analysis (LPC) to estimate the vocal tract. However, in Aparat the default autoregressive model in use is a discrete all-pole (DAP) model [9], as LPC and minimum variance distortionless response (MVDR) model [10] can be found on advanced settings panel.

Former GIF methods in Aparat, removed from an updated version 2.0:

The two deleted methods are direct inverse filtering (DIF) method (simpler version of an IAIF method) and experimental inverse filtering method. The DIF method is basically a simple version of IAIF method, as the DIF method follows the same filtering process as IAIF but finishes after the first integration step [1]. The experimental method was never fully finished and it did not run properly.

2 Installing Aparat

Aparat is a Matlab-based programme that can be run with or without MATLAB software. The installation instructions for both cases are explained.

2.1 When having MATLAB licence

For MATLAB users installing Aparat is really straightforward: MATLAB users' installation package is selected and uncompressed to a suitable location (for example, under the users personal MATLAB working directory). Aparat is opened by invoking its main function "aparat.m".

Aparat uses Matsig, an object-oriented signal processing library for MATLAB. Previous releases of TKK Aparat required Matsig to be separately installed, but the current package includes also Matsig.

2.2 Without MATLAB

Installation and the usage of the stand-alone version of Aparat requires the Matlab Component Runtime to be downloaded and installed. It is, however, available together with Aparat files on the website of the author.

The Matlab Component Runtime can be installed by double-clicking the file and following the instructions on the screen.

Currently, no installation program exists for Aalto Aparat. Instead, it can be installed by following the steps below:

1. Load the right installation package according to your operating system.
2. Create a folder for the programme
3. Uncompress all files from the installation package into that folder.
4. (Optional) Create a shortcut to Aparat.exe on the desktop by right-clicking the desktop, selecting New -> Shortcut, and then selecting Aparat.exe.

Now, Aalto Aparat may be run by double-clicking the Aparat.exe file.

3 Work with Aparat

3.1 File types

Aparat processes .wav files and save the processed data to .wav, .mat, and .tab files. The .wav files in process may be in stereo or mono, and the optimal length of the sample is from 0.5 to 10 seconds.

The obtained waveforms can be saved to .wav or .mat files. The saving process is presented in detail in the section Saving data.

3.2 The main windows and their features

Aalto Aparat has two main windows: control window on the left and signal view window on the right (Fig. 2) . They both appear automatically when opening the programme.

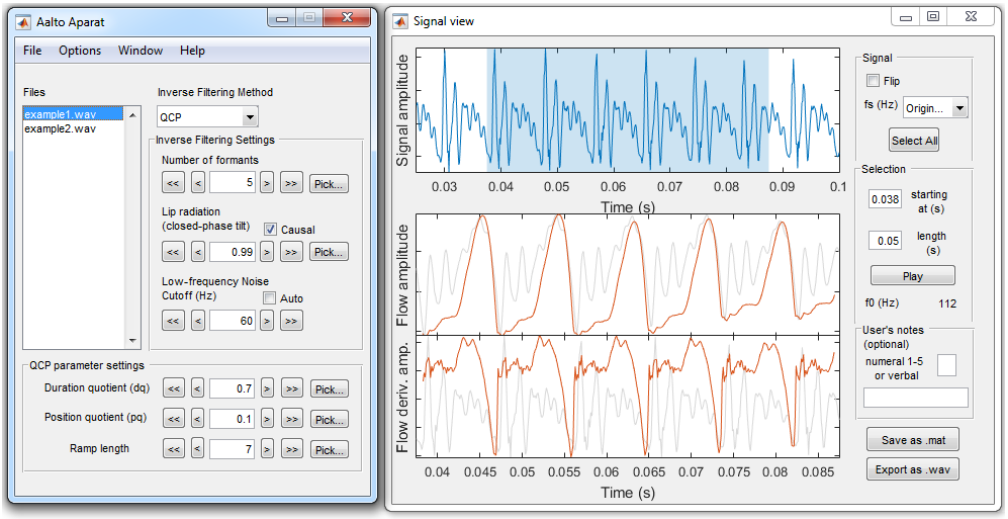


Figure 2: Aalto Aparat main windows: Control window on the left and signal view on the right

In Aparat control window there is a list of all the .wav files that exist in the same folder as the programme (Fig. 3). If some of the files also has a corresponding .mat file, there is an asterisk (*) in the beginning of the file name. More about files and saving: Saving data.

Despite different operating systems Aparat should always look the same, with slightly varying window frames.

Aparat can be closed by clicking the close icon on the top right corner of the control window or via File menu -> Exit.

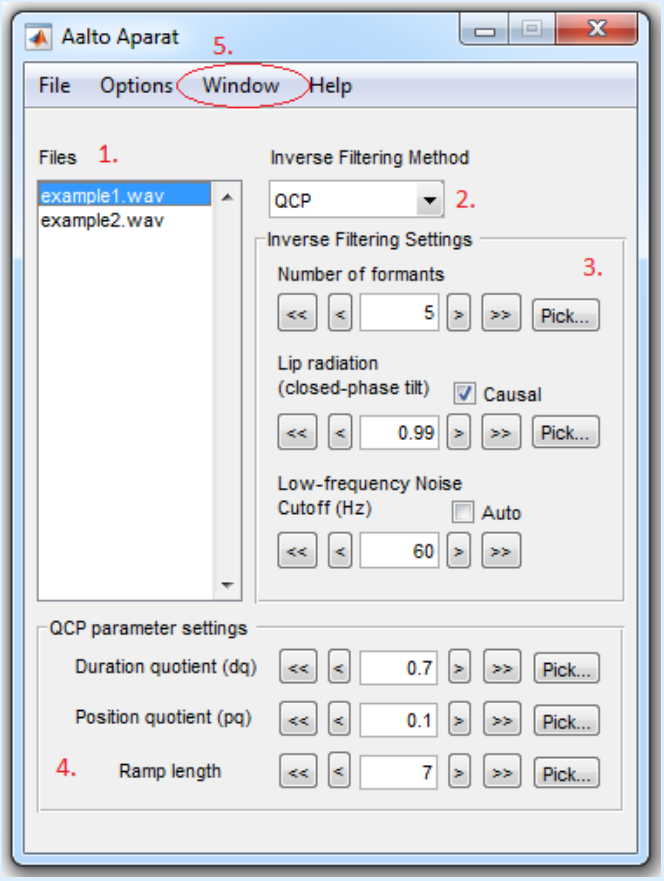


Figure 3: Close-up to control window

Legend
1 List of .wav files in the programme folder
2 Glottal inverse filtering (GIF) method popup-selection
3 General filtering settings for all GIF methods
4 Method dependent parameter settings
5 Special windows

6

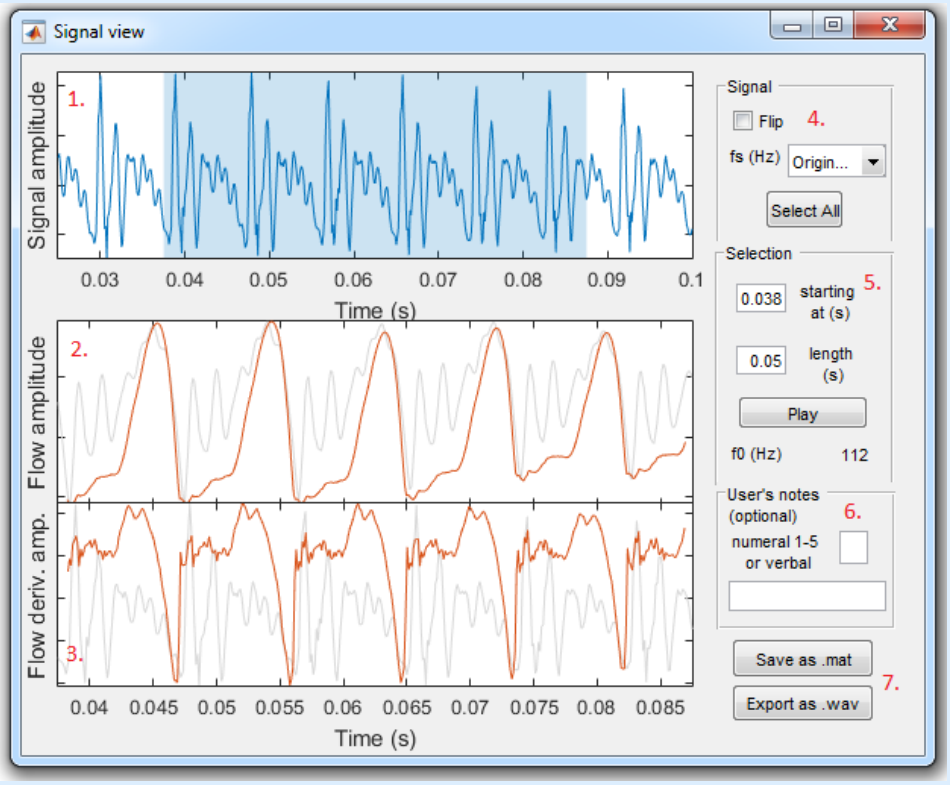


Figure 4: Close-up to signal view

Legend
1 Original signal and selection
2 Glottal flow
3 Glottal flow derivative
4 Flipping signal and sampling frequency selection (popup)
5 Selection starting point and length
6 Optional notes (numeral and verbal)
7 Save buttons

3.3 Parameter settings

The user is able to manipulate a wide selection of parameters in updated Aparat. Not all filtering methods use exactly the same parameters but the logic is the same behind all the similar handles.

The handles of the following three general parameters are always visible on the control window: the number of formants, lip radiation coefficient, and low-frequency noise cutoff (in Hz) (Fig. 5). Additionally, there are QCP parameter and IAIF parameter panels under the filtering setting panel: they will become visible when corresponding filtering method is selected (Fig. 6). The QCP method has three and IAIF two extra parametes. All the parameters and their influence on the filtering process can be seen on a Table 1.

Table 1: The GIF parameter settings on the Aparat control window

Parameter	Effect	Valid values	Other
Number of formants	Defines the order of the vocal tract filter	2 - 30	
Lip radiation effect	Affect on the integration coefficient of the flow waveform	0.9 - 1	
Low-frequency noise cutoff (Hz)	(filter)	depends on f0	
Duration quotient (DQ)	Relative length of the non-attenuated section of AME window	0.1 - 1	For QCP only
Position quotient (PQ)	Relative starting point of the non-attenuated section of AME window	0 - 0.15	For QCP only
Ramp length	Number of transitional samples of the AME window	1 - 20	For QCP only
AR modeling method	Selection of the vocal tract modelling method	DAP, LPC, and MVDR	For IAIF only
Glottis filter order	Prediction order of the glottal flow filter	2 - 8	For IAIF only

8

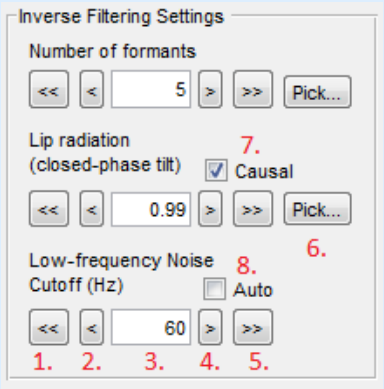


Figure 5: Close-up to general parameter settings

Legend
1 Decrease the value by two steps
2 Decrease the value by one step
3 Open box to show the value and edit it manually
4 Increase the value by one step
5 Increase the value by two steps
6 Pick button (see more at Fig. 7)
7 Causality option (on / off)
8 Set low-freq noise cut automatically (selecting "auto" disables other low-freq handles)

The best parameter values can be found by iterating step by step or writing the exact numbers in the open box. However, as this is not always the fastest way to work there is also a "Pick" button in Aparat. "Pick" allows the user to hand-pick the best-looking waveform.

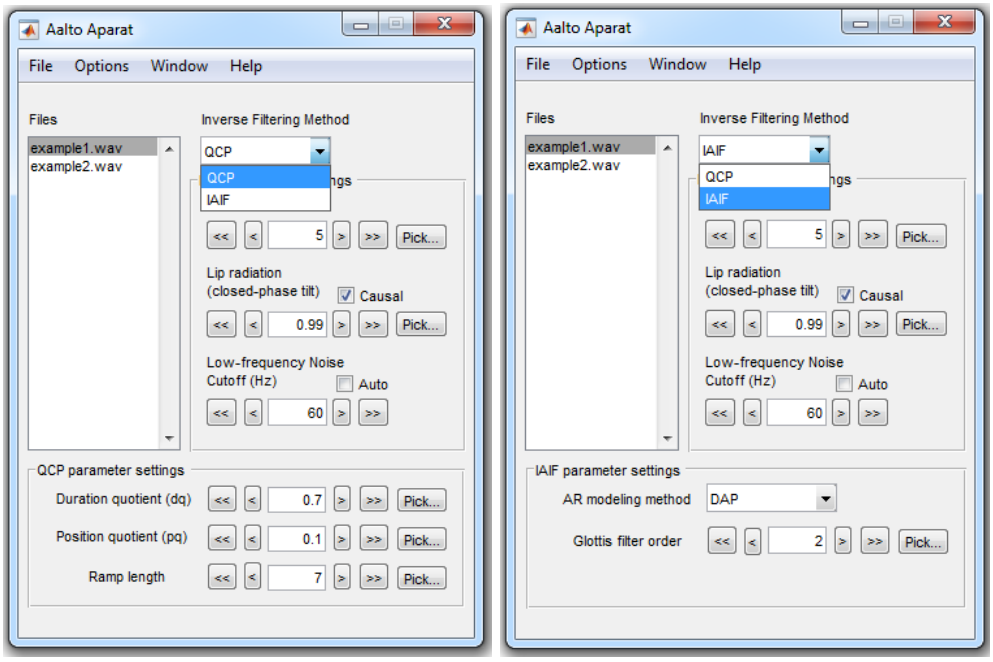


Figure 6: The appearance of the control window changes according to method in use

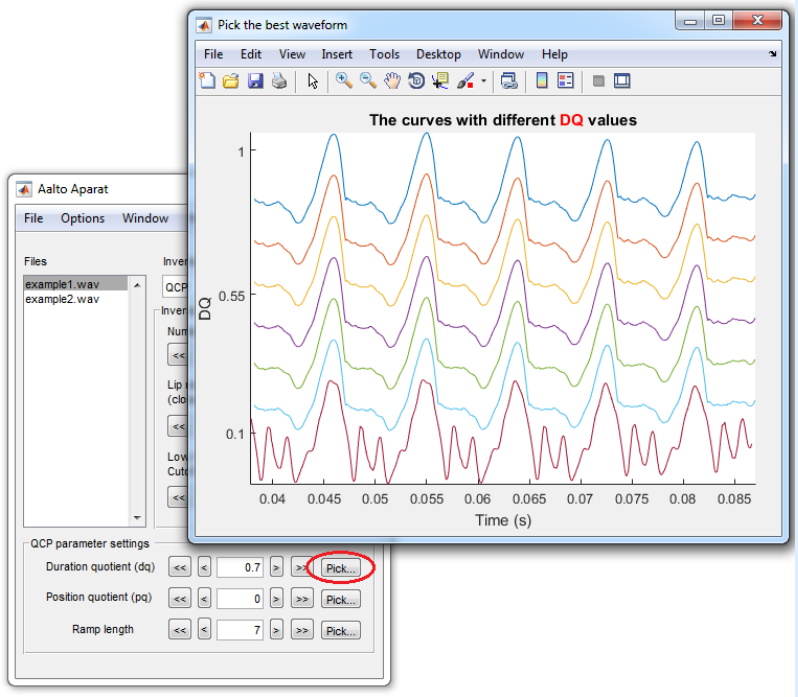


Figure 7: "Pick the best waveform" button

Pick button opens a new window and visualizes the range of different waveforms on all the possible values of the parameter in question (see page 1). In this example case the parameter is duration quotient (DQ) of the QCP method: as all DQ values must be between 0.1 and 1, the bottom waveform is drawn with PQ value 0.1 and the top with a value 1, leaving the rest of the scale in between. The number of waveforms drawn varies slightly between different parameters (from 5 to 7 individual waveforms). The best waveform can be chosen by clicking it with a cursor.

3.4 Other options and tools

In addition to parameter settings Aparat has several different tools to analyse the inverse filtered signal. These tools can be found on a menu bar of the control window (see Fig. 8).

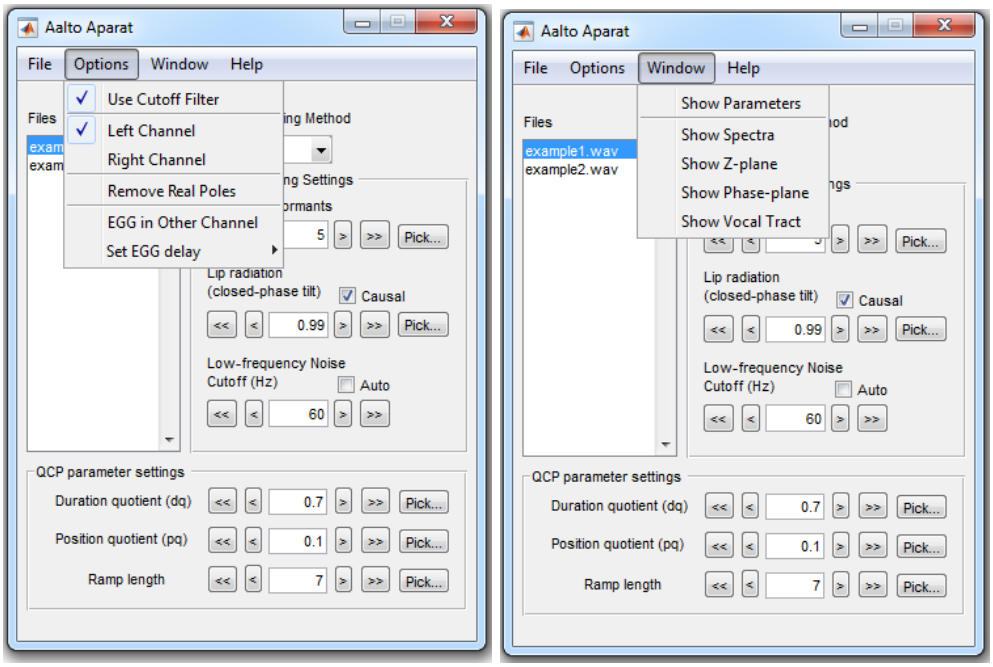


Figure 8: Advanced options and tools can be found on the menu bar

In **Options** menu there are special options in connection to the input signal: channel election, cutoff filter on/off, real poles removal from the vocal tract model, and settings for EGG analysis. In **Window** menu the glottal source parametrisation tool and graphic presentations of spectra, z-plane, phase-plane, and vocal tract can be viewed (Fig. 9-13).

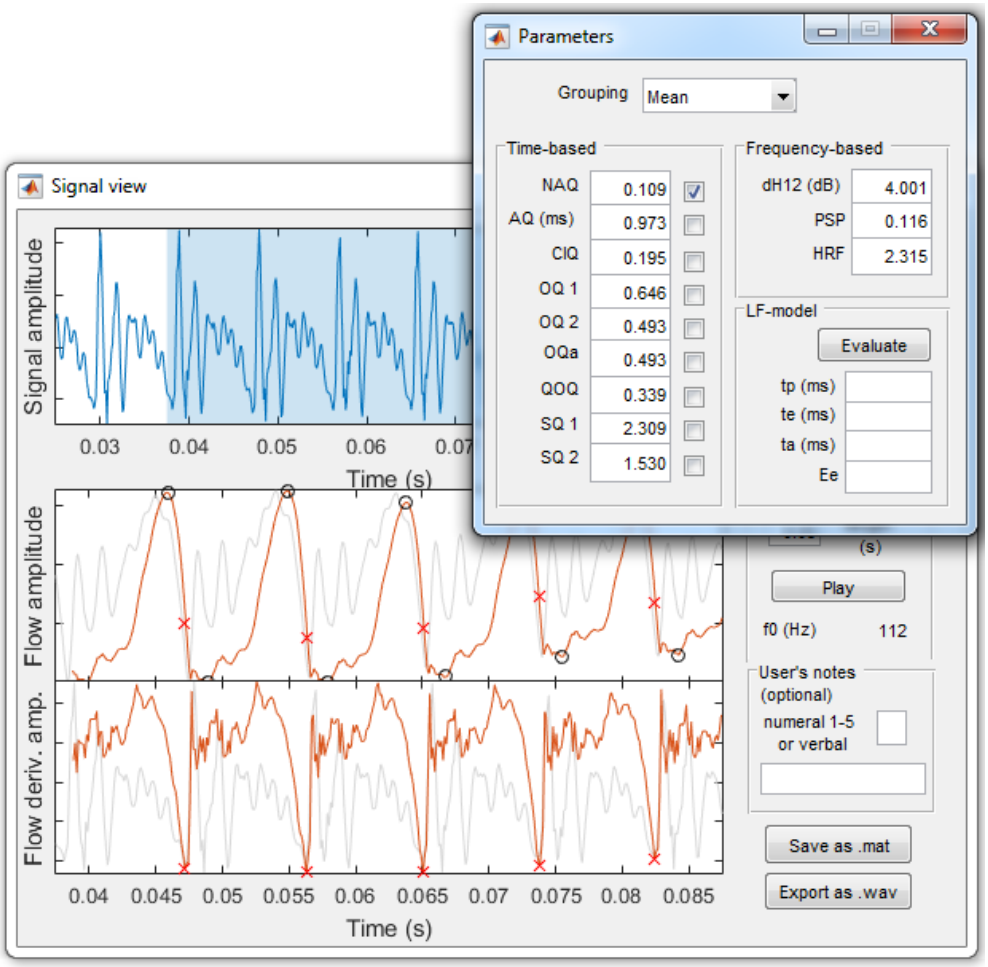


Figure 9: The parameter window shows the parameters computed from a glottal flow estimate. By clicking the checkbox next to the value (i.e. NAQ as in the figure) the time-domain instants related to the computation of the parameter can be seen on the signal view window [1].

13

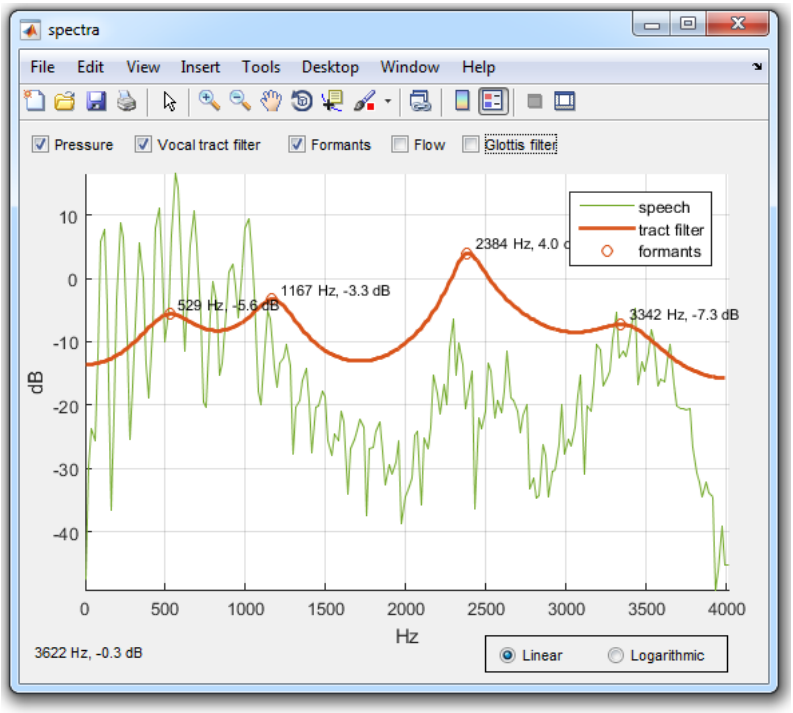


Figure 10: The spectra window is able to illustrate signal pressure, vocal tract filter, formants, glottal flow and glottis filter all in the same plot, using either the linear or logarithmic scale. In this example, only pressure, vocal tract filter and formants are selected: the selection is made by clicking corresponding check boxes.

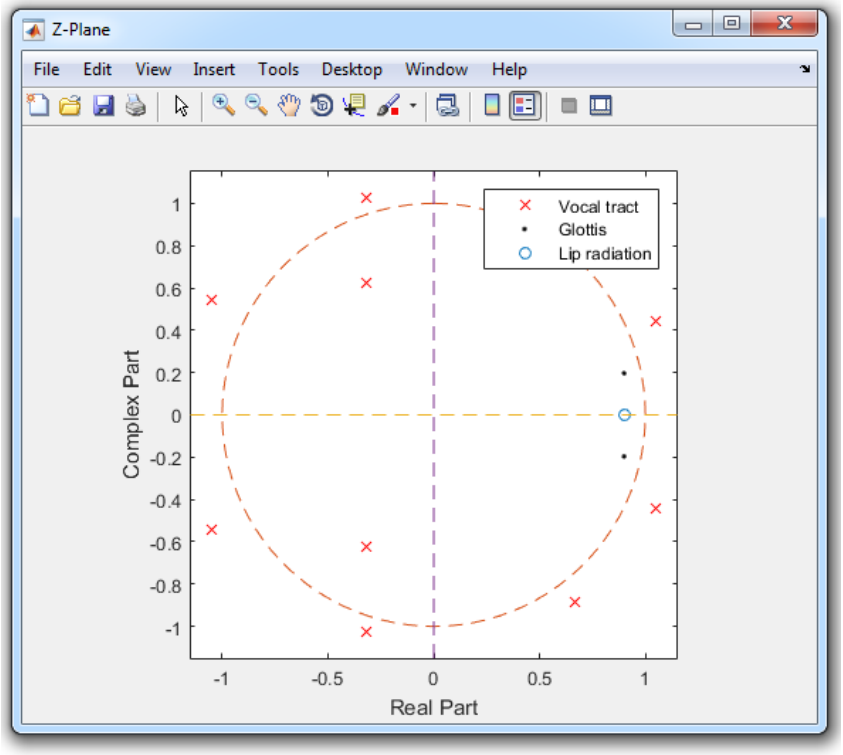


Figure 11: Z-plane figure plots the current vocal tract filter estimate on the z-plane. The resonances of different stages are indicated differently.

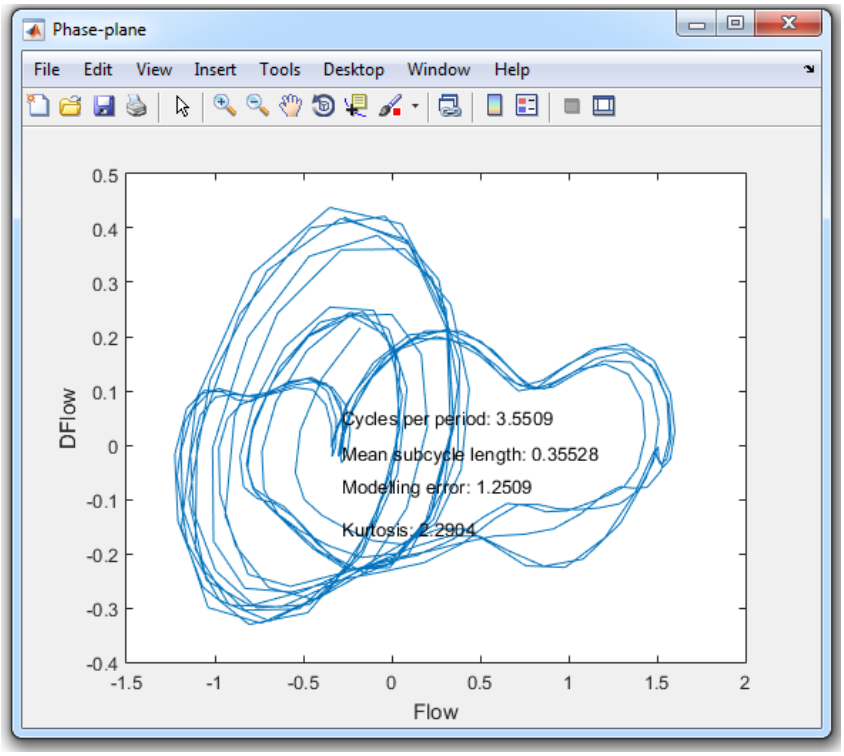


Figure 12: Phase-plane figure shows an xy-plot with the glottal flow samples on the x-axis and the corresponding samples of the flow derivative on the y-axis. It also announces the cycles per period, mean sub-cycle length, modelling error and kurtosis.

16

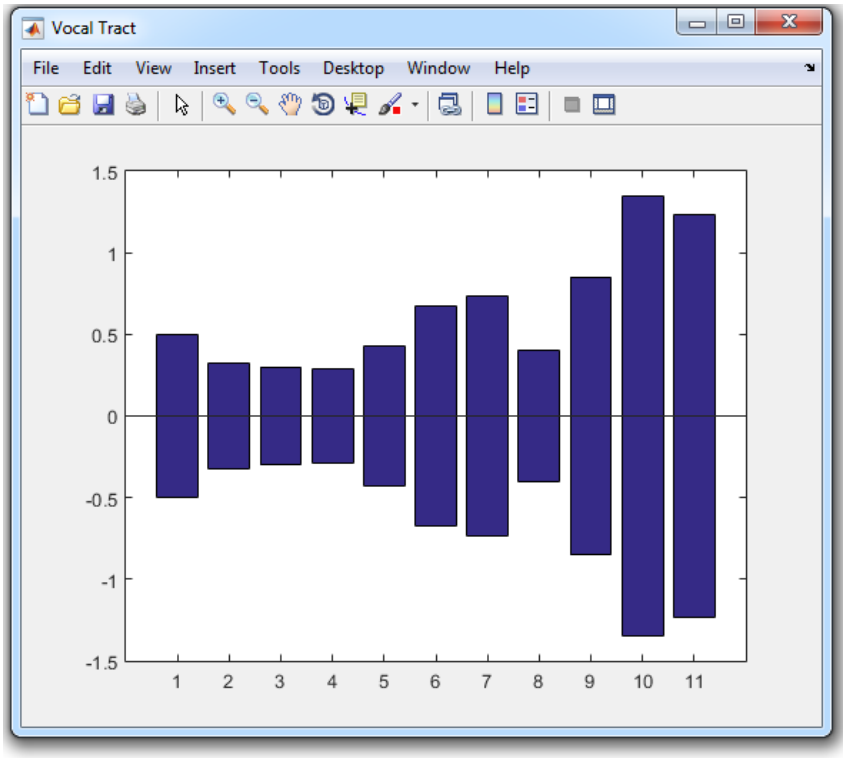


Figure 13: Vocal tract view shows a plot of the cross-sectional diameter of the tube model derived from the vocal tract filter.

3.5 Saving data

Aparat provides two ways to save your work: saving all current parameter values and selection as a .mat file and saving the selected sound pressure and glottal flow as .wav files. The save buttons can be found on both the control window (under the File menu, see Fig. 14) and the bottom right corner of the signal view window.

The **Export as .mat** command names the new .mat file after the original wav. file in process, saves all current values and places the file itself to the work space folder of the programme. The existence of the .mat file can be viewed on the Aparat control window as there is an asterisk (*) in the beginning of the name of the original signal (Fig. 3), and the saved parameter values are now default settings for this particular .wav file . By contrast, the **Save as .wav** command creates a new folder (exportWavs) and saves the samples there, and they are not shown in the Aparat control window without changing the work space folder. The names of the .wav samples include the name of the original signal followed by "Flow" (for glottal flow) or "Pressure" (for selected sound pressure).

Note: to make several different samples of the same signal there must be several differently named copies of the original as well, as Aparat overwrites the existing sample(s) with a new one if the names are identical.

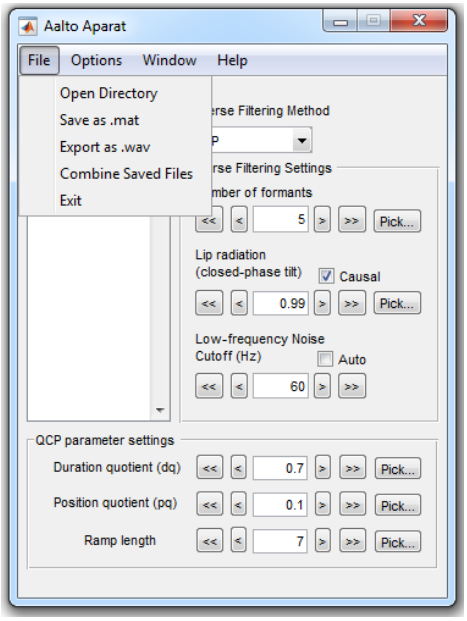


Figure 14: Saving options as well as exit can be found under File menu

After making various .mat files there may be need to compare them and process them as a one file. For that purpose there is a **Combine Saved Files** command on the File menu (Fig. 14). The command collects all the .mat files from the work space and packs them to a file named "combined.tab".

References

[1] M. Airas. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008.

[2] M. Airaksinen, T. Raitio, B. Story, and P. Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE Transactions in Audio, Speech and Language processing*, 22(3):596–607, 2014.

[3] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

[4] D. Wong, J. Markel, and A. Gray Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *Audio, Speech, and Language Processing, IEEE Transactions on*, 27(4):350–355, 1979.

[5] C. Ma, Y. Kamp, and L. Willems. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1):66–81, 1993.

[6] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical society of America*, 134(2):1295–1313, 2013.

[7] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):994–1006, 2012.

[8] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):1–13, 1985.

[9] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *Signal processing, IEEE Transactions on*, 39(2):411–423, 1991.

[10] M. Wölfel and J. McDonough. *Front Matter, in Distant Speech Recognition*. John Wiley & Sons, Ltd, 2009.

B Finnish text for the recording sessions

The following instructions and text paragraphs were given to the subjects at the recording session. The text part contains phrases typically used in weather forecasts.

Lue alla oleva sääilmaisuja sisältävä teksti normaalilla puhenopeudella läpi kolme kertaa. Pienet virheet puheessa eivät haittaa, mutta jos kesken kaiken alkaa esim. kovasti naurattaa, voit aloittaa kappaleen alusta.

”Saatujen tietojen mukaan päättynyt heinäkuu oli jo kolmas peräkkäinen hyvin lämmin heinäkuu. Kuukauden keskilämpötilat lähestyivät lämpöennätyksiä, jotka on mitattu tuhatyhdeksänsataaluvun alussa. Heinäkuun viimeinen päivä oli äärimmäisen harvinainen, koska hellettä oli kaikkialla Suomessa.

Kaatosadetta saatiin eilen illalla Oravaisten kunnassa. Sademäärä oli noin satavisikymmentä millimetriä. Vuodenvaihdetta juhlitaan talvisäässä. Ilmatieteenlaitoksen maanantaina laatiman ennusteen mukaan vuosi vaihtuu talvisessa pakkassäässä. Voimakas korkeapaine liikkuu maanantaina Suomen pohjoisosan yli kaakkoon. Jääkartat ovat nyt päivittäin saatavilla laitoksen verkkosivuilla.

Kaasukehiä sekä Auringon vaikutusta kaasukehiin tutkitaan, samoin ilmanlaadun ja ilmansaasteiden vaikutuksista terveyteen ja ympäristöön. Tutkimuksen avulla saadaan päätöksentekijöille tietoa ilmanlaadusta.”