Vocal Aging Explained by Vocal Tract Modelling: 2008 JHU Summer Workshop Final Report[1]

Peter Beyerlein (peter.beyerlein@tfh-wildau.de)
Andrew Cassidy (acassidy@jhu.edu)
Varada Kholhatkar (kolha002@d.umn.edu)
Eva Lasarcyk (evaly@CoLi.Uni-SB.DE)
Elmar Nöth (noeth@informatik.uni-erlangen.de)
Blaise Potard (potard@loria.fr)
Stephen Shum (sshum@berkeley.edu)
Young Chol Song (nskystars@gmail.com)
Werner Spiegl (spiegl@informatik.uni-erlangen.de)
Georg Stemmer (georg_stemmer@web.de)
Puyang Xu (puyangxu@gmail.com)

[1]We gratefully acknowledge a number of collaborators that contributed to this project before and during the workshop: Andreas Andreou, Peter Birkholz, Allen Gorin, James Harnsberger, Jonathan Harrington and Sanjeev Khudanpur.
The authors appreciate the support of the JHU CLSP 2008 summer workshop and its sponsors for enabling this research.

Contents

Chapter 1

Introduction

A person's voice can change due to many reasons, including aging, or being in different emotional and pathological conditions. Physiologically, when a person ages, their vocal tract lengthens, pulmonary functions reduce, the laryngeal cartilages ossify, stiffness in the vocal fold increases, and closures are reduced [Leeu 04].

On the acoustical side, the fundamental frequency F_0 is known to increase in males and decrease in females, with variability increasing with age [Linv 02]. People tend to speak slower, and aged people have more noise in their voices [Ferr 02]. Formant frequencies are generally known to lower with aging, though these results are not always consistent. These findings show that there are features in speech that correspond with age.

It was the purpose of this workshop to look into the speech signal and features calculated from it and identify changes that can be attributed to aging. We decided to not only look at standard features for automated speech processing (e.g. Mel Frequency Cepstrum Coefficients, MFCCs), but also prosodic features and implement models, that invert the articulatory process. Thus we can deduce parameters for the excitation signal and the vocal tract. The degree of change in the standard, the prosodic and these phonetically motivated features is evaluated with respect to the question how good these features can predict the age of the speaker.

For our research we used both longitudinal data (one speaker over a long age period) and cross-sectional data (many speaker of different age groups). When trying to look at the aging process it would be best to keep all other factors (speaker, communication situation, emotion of the speaker, recording devices, ...) constant. Of course it is very difficult to get recordings of one speaker over a long period and impossible to keep the recording devices constant over this period. The longitudinal data that were available to us were from the same communication situation, albeit certain factors like routine and experience in public speech change over time and to a much higher degree the microphones and the other recording devices. It was our expectation that articulatory features would encode the channel information to a much lesser degree than standard MFCC features. We decided to verify findings from the longitudinal data with cross-sectional data, where many speakers from different age groups spoke the same text under identical recording conditions. To our knowledge this was the first research where changes in the aging voice were examined for these two different kinds of data.

For longitudinal data ([Harr 06]) we chose

- **The Queen of England Christmas Speeches:**

  The radio broadcast Christmas speeches of the Queen of England from 1952 - 2002 (some years missing, a total of 30 speeches) were digitized by BBC with different sampling rates and # of bits/sample. We converted the files to 16 kHz & 16 bits/sample. The speeches were about 5 minutes in length, totaling about 2.5 hours of speech. There was a strong channel effect in these recordings, because of the difference of recording quality and microphone over the years. This effect was also detected during human evaluation of the data.

- **Alistair Cooke's Letter from America Broadcast Data:**

  This data consists of 30 recordings of Alistair Cooke's *Letter from America* broadcasts, each under 25 minutes of speech, with a total of 10.6 hours in length. Recordings ranged from 1947 to 2003. These recordings also suffered from a change in channel over the years. There was also a language effect in the recordings; although Alistair Cooke initially had a strong British accent, he gradually received more influence from American English over the years.

For cross-sectional data ([Harn 08]) we chose the

- **University of Florida - Vocal Aging Database (UF-VAD):**

  The UF-VAD database consists of a Rainbow Passage, Grandfather Passage, sustained vowels, and 16 SPIN-sentences from 75 male and 75 female American English speakers. The speakers from each sex group were evenly divided into 25 young, 25 middle-aged and 25 old with groups ranging from 18-29, 40-55 and 62-92 years of age, respectively. The recordings were made between 2003 and 2007, and were recorded with the same microphone with the same text, minimizing channel and language effects. Each speaker spoke each of the passages, vowels and sentences once, which amounted to about two minutes of speech, totaling the entire database to about 5 hours in length. Figure 1.1 shows the characteristics of this database.

The rest of this report is organized as follows: In chapter 2 we describe the implementation of the glottal excitation model that we used in order to extract excitation parameters from the speech signal. For the parameters derived from the vocal tract inversion (described in chapter 4) the first three formants have to be extracted from the speech signal. To do this we used two well known and freely available formant trackers. To exclude possible changes in formant values based on changing recording conditions rather in aging changes we analyzed the errors of the algorithm for some of the queens data at different ages. These analyzes are described in chapter 3. The MFCCs and the prosodic parameters are described in chapter 5. Experiments to predict the age of the speaker based on different feature groups are treated in chapter 6. The report finishes with a summary.

Talker Group		Chronologic Age			Perceived Age		
		Mean	Min	Max	Mean	Min	Max
Young	Male	**22**	18	29	**28**	21	37
	Female	**20**	18	24	**22**	17	29
Middle	Male	**49**	41	55	**43**	31	57
	Female	**48**	40	55	**43**	34	58
Older	Male	**78**	62	92	**62**	49	73
	Female	**79**	65	89	**65**	42	82

Figure 1.1: Characteristics of the UF-VAD database

Chapter 2

Glottal Excitation Optimization

2.1 Introduction

2.1.1 Motivation

The goal of this work is to investigate the influence of age-related changes of the larynx on speech. Harnsberger et al. [Harn 08] gave a recent literature overview on these changes, including (i) an increased stiffening of the vocal folds, (ii) laryngeal cartilage ossification, and (iii) a reduction in vocal fold closure. Unfortunately, the relationship of the physiological changes with the speech signal produced by the aging speaker is still unclear. Linville [Linv 96] has described a number of voice features that listeners often considered to be characteristic for aged speakers. Many of them, like lower vocal pitch, increased harshness or hoarseness, increased strain, higher incidence of voice breaks, vocal tremor and increased breathiness, may be related to physiological changes of the larynx. Only very few researchers, however, have been able to successfully map perceptual features of voice quality to objective quantities measurable from recorded speech. An exception is the lowered pitch, as several sources in the literature observed a relationship between the fundamental frequency and age [Harn 08]. For male speakers there is an increase of the mean fundamental frequency with age; for older female speakers there is either no change at all or a decrease.

This work is intended to be a first step in establishing a relationship between the observed voice quality and changes of the larynx. The goal is to be able to describe the age-related changes of the voice quality by the parameters of a physical model of the glottis. The features derived from this physical model should represent a speaker's age better than conventional methods. Ultimately, these features could lead to age normalization methods.

As the relationship between the pitch and age is already known and can be extracted from the speech signal without an elaborate model, we exclude the influence of the pitch as much as possible. Instead, we concentrate on all other aspects of the speech signal, such as increased strain, harshness, hoarseness, or creakyness.

2.1.2 Approach

Our approach is based on the source-filter model of the speech generation process. Air flows from the lungs through the vocal tract, consisting of laryngeal cavity, pharynx, oral and

nasal cavity. Voiced speech sounds are generated by a glottal source signal, also called an *excitation signal*, that is filtered by the vocal tract. In order to be able to separate the influence of glottis and vocal tract, it is assumed that the vocal tract can be represented by a *linear filter*. As shown in Fig. 2.1, linear prediction is applied to obtain the vocal tract configuration for each time frame. An approximation of the excitation signal is the LPC

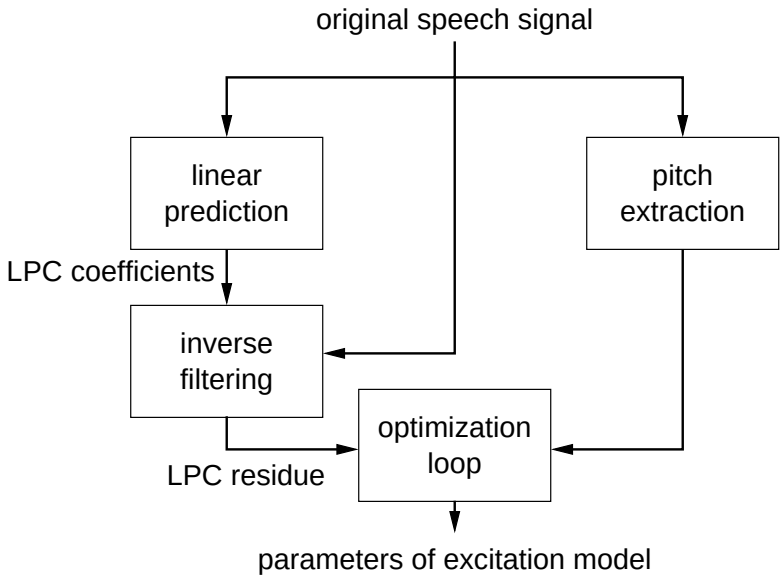

Figure 2.1: Optimization procedure for a single time frame of voiced speech.

(Linear Predictive Coding) residue, which is computed by inverse filtering the speech signal with the LPC filter. We use a parametric physical model of the vocal folds to represent the excitation signal. As it is impossible to derive the true model parameters analytically, a data-driven optimization procedure aims at fitting the synthetic excitation signal generated by the model as close as possible to the LPC residue and the estimated pitch. This optimization procedure is denoted as the *optimization loop* in Fig. 2.1. We assume that the parameters corresponding to the closest fit represent the true excitation signal adequately. The final parameters of the excitation model are then analyzed with regard to age correlation.

2.1.3 Overview

The next section introduces the employed excitation model, which is a two-mass vocal fold model introduced by Stevens in [Stev 98]. In Sec. 2.3 the parameter optimization loop using the simplex and simulated annealing algorithms is described, followed by the experiments and results in Sec. 2.4. We conclude in Sec. 2.5 with a short summary and an outlook on future work.

2.2 Excitation Model

2.2.1 Two Mass Model

Our approach estimates the parameters of a physical glottis model from speech data that has been recorded at different ages of a speaker. The goal is to find age-related changes in the model parameters. Therefore the glottis model used should ideally have physically meaningful parameters, in contrast to just describing the shape of the excitation signal. In order to be able to optimize the parameters using a limited amount of data, their number should be as small as possible. As the same time the model should be flexible enough to adequately represent age-related changes of the voice quality.

Considering the above mentioned requirements we employed the two-mass vocal fold model introduced by Stevens [Stev 98] and illustrated in Fig. 2.2.

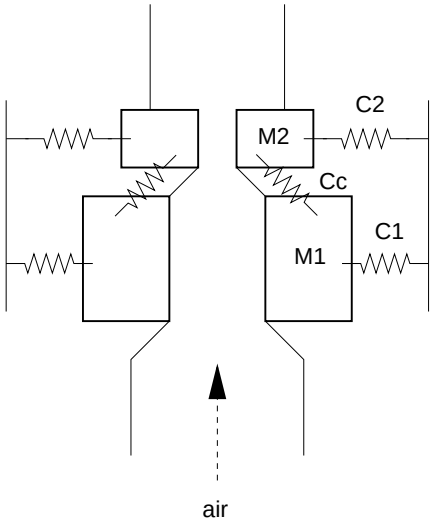

Figure 2.2: Two-mass vocal fold model by Stevens [Stev 98].

The symmetrical model consists of two pairs of masses, larger ones representing the lower part of the vocal folds, and small ones representing the upper part of the vocal folds. The mass parameters are M_1 and M_2 as shown in Fig. 2.2. The masses each move on springs and are connected together by an additional spring. The compliances of the springs are determined by the parameters C_1, C_2 and C_c (for the spring that connects M_1 with M_2). Note that parameters for the masses and compliances are given as *mass per unit length* and *compliance per unit length*, i.e. they may change when the vocal folds are stretched. Air flows from bottom to top through the glottis when both M_1 and M_2 have a positive displacement, as shown in Fig. 2.2.

The excitation function for the two-mass vocal fold model by Stevens is obtained in three steps. First, the displacements $x_1(t)$ and $x_2(t)$ of the lower and upper part of the vocal folds over time t are computed. The width of the glottal opening $d(t)$ is defined to be $\min(x_1(t), x_2(t))$. Second, from the width of the opening, the airflow $U_g(t)$ through the glottis is determined. In the third step, taking the derivative of $U_g(t)$ results in the excitation function.

The whole process of the excitation function computation is described in Chapter 2 of [Stev 98]. However, since many details are omitted in [Stev 98], we re-derived the equations on our own and also had to consider simplifications. Consequently the next paragraphs describe the excitation function computation as it was implemented for the experiments reported in Sec. 2.4. The initial values for all parameters are taken from [Stev 98] as well, however, some parameters cannot be found in the book. Therefore Tab. 2.1 lists all the default parameter settings that we used for the vocal fold model.

Table 2.1: Default Parameter Settings for the Vocal Fold Model

		male	female
M_1	[gm/cm]	0.1	0.04
M_2	[gm/cm]	0.02	0.008
C_1	[cm^2 / dyne]	$3e^{-5}$	$1.9e^{-5}$
C_c	[cm^2 / dyne]	$5e^{-5}$	$3.1e^{-5}$
d_1	[cm]	0.2	0.133
x_0	[cm]	0.01	0.005
l	[cm]	1.0	0.7
P_s	[dyne / cm^2]	8000	
ΔP_g	[cm H2O]	8.0	
$12\mu h$		0.00001	
ϕ		0.0	
ρ	[gm/ cm^3]	$1.14e^{-3}$	
κ		1.0	
ζ		0.7	

Glottal Displacement

The first step is to compute the displacements $x_1(t)$ and $x_2(t)$ of the lower and upper part of the vocal folds over time t. There are several critical points which mark the different stages of the movement of the masses. The models dynamics are described in a piecewise fashion for each time segment. The first segment ranges from $t_0 = 0$, when mass M_1 starts to move, to t_1, when mass M_2 starts to move as well.

Displacement of the lower part of the vocal folds from t_0 to t_1. According to Stevens [Stev 98] the displacement $x_1(t)$ for the segment from t_0 to t_1 is described by the differential equation:

$$M_1 x_1'' + \frac{1}{C_1}(x_1 - x_0) = P_s d_1 \tag{2.1}$$

where x_0 is a constant that stands for the resting position of M_1 in the absence of any force, P_s is the subglottal pressure, and d_1 is the average vertical length of the lower portion of the

vocal fold. Rearranging leads to

$$C_1 M_1 x_1'' + x_1 = (C_1 P_s d_1 + x_0) \tag{2.2}$$

The general solution to this differential equation is:

$$\begin{aligned} x_1(t) &= \sin(\omega_1 t)k_2 + \cos(\omega_1 t)k_1 + (C_1 P_s d_1 + x_0) && (2.3)\\ x_1'(t) &= \cos(\omega_1 t)k_2 - \sin(\omega_1 t)k_1 && (2.4) \end{aligned}$$

where the natural frequency ω_1 of mass M_1 is $\omega_1 = 1/\sqrt{C_1 M_1}$ and k_1, k_2 are constants determined by boundary conditions. Applying the initial conditions $x_1(t_0) = 0$, and $x_1'(t_0) = 0$, we solve for the constants k_1, k_2:

$$\begin{aligned} x_1(0) = 0 &= \sin(0)k_2 + \cos(0)k_1 + (C_1 P_s d_1 + x_0) && (2.5)\\ 0 &= 0 + k_1 + (C_1 P_s d_1 + x_0) && (2.6)\\ k_1 &= -(C_1 P_s d_1 + x_0) && (2.7)\\ x_1'(0) = 0 &= \cos(0)k_2 - \sin(0)k_1 && (2.8)\\ 0 &= k_2 - 0 && (2.9)\\ k_2 &= 0 && (2.10) \end{aligned}$$

Submitting back into equations 2.3 and 2.4 results in:

$$\begin{aligned} x_1(t) &= [\cos(\omega_1 t)][-(C_1 P_s d_1 + x_0)] + (C_1 P_s d_1 + x_0) && (2.11)\\ &= (C_1 P_s d_1 + x_0)[1 - \cos(\omega_1 t)] && (2.12)\\ x_1'(t) &= -\sin(\omega_1 t)[-(C_1 P_s d_1 + x_0)] && (2.13)\\ &= (C_1 P_s d_1 + x_0)\sin(\omega_1 t) && (2.14) \end{aligned}$$

Displacement of the lower part of the vocal folds for $t > t_1$. At point t_1, the upper mass M_2 starts to move and both parts of the vocal folds are separated. The model does not specify t_1, therefore we select it to be at the point in time when $x_1(t)$ has reached half of the maximum displacement. That is, t_1 is the quarter period point of $x_1(t)$. The displacement of M_1 at t_1, i.e. $x_1(t_1)$ is referred to as x_{10}:

$$\begin{aligned} t_1 &= \frac{1}{2}\pi \cdot \frac{1}{u_0} = \sqrt{M_1 \cdot C_1} \cdot \frac{1}{2}\pi && (2.15)\\ x_{10} &= P_s \cdot d_1 \cdot C_1 + x_0 && (2.16) \end{aligned}$$

The opening of the vocal folds at t_1 results in a drop of the subglottal pressure P_s to 0, leading to an altered movement of M_1. In order to compute $x_1(t)$ for $t > t_1$, we have to take into account $P_s = 0$ when determining k_1 and k_2. We select the following boundary conditions to ensure a smooth transition between the two segments of $x_1(t)$ at t_1: $x(t_1) = x_{t1}$ and $x'(t_1) = dx_{t1}$, where x_{t1} and dx_{t1} are the displacement and velocity at t_1, found using Eq. 2.12 and Eq. 2.14.

$$\begin{aligned} x_1(t_1) = x_{t1} &= \sin(\omega_1 t_1)k_2 + \cos(\omega_1 t_1)k_1 + (C_1 P_s d_1 + x_0) && (2.17)\\ x_1'(t_1) = dx_{t1} &= \cos(\omega_1 t_1)k_2 - \sin(\omega_1 t_1)k_1 && (2.18)\\ \sin(\omega_1 t_1)k_1 &= \cos(\omega_1 t_1)k_2 - dx_{t1} && (2.19)\\ k_1 &= \frac{\cos(\omega_1 t_1)k_2 - dx_{t1}}{\sin(\omega_1 t_1)} && (2.20) \end{aligned}$$

Inserting $P_s = 0$ and solving for k_2:

$$\begin{aligned}
x_{t1} &= \sin(\omega_1 t_1)k_2 + \cos(\omega_1 t_1)k_1 + (0 + x_0) && (2.21)\\
&= \sin(\omega_1 t_1)k_2 + \cos(\omega_1 t_1)\left(\frac{\cos(\omega_1 t_1)k_2 - dx_{t1}}{\sin(\omega_1 t_1)}\right) + x_0 && (2.22)\\
&= \sin(\omega_1 t_1)k_2 + \frac{\cos^2(\omega_1 t_1)k_2}{\sin(\omega_1 t_1} - \frac{dx_{t1}\cos(\omega_1 t_1)}{\sin(\omega_1 t_1)} + x_0 && (2.23)\\
\sin(\omega_1 t_1)x_{t1} &= \sin^2(\omega_1 t_1)k_2 + \cos^2(\omega_1 t_1)k_2 - dx_{t1}\cos(\omega_1 t_1) + \sin(\omega_1 t_1)x_0 && (2.24)\\
x_{t1}\sin(\omega_1 t_1) &= [\sin^2(\omega_1 t_1) + \cos^2(\omega_1 t_1)]k_2 - dx_{t1}\cos(\omega_1 t_1) + x_0\sin(\omega_1 t_1) && (2.25)\\
x_{t1}\sin(\omega_1 t_1) &= k_2 - dx_{t1}\cos(\omega_1 t_1) + x_0\sin(\omega_1 t_1) && (2.26)\\
k_2 &= x_{t1}\sin(\omega_1 t_1) + dx_{t1}\cos(\omega_1 t_1) - x_0\sin(\omega_1 t_1) && (2.27)\\
k_2 &= (x_{t1} - x_0)\sin(\omega_1 t_1) + dx_{t1}\cos(\omega_1 t_1) && (2.28)
\end{aligned}$$

Substituting back in to find k_1:

$$\begin{aligned}
k_1 &= \frac{\cos(\omega_1 t_1)k_2}{\sin(\omega_1 t_1)} - \frac{dx_{t1}}{\sin(\omega_1 t_1)} && (2.29)\\
&= \frac{\cos(\omega_1 t_1)[(x_{t1} - x_0)\sin(\omega_1 t_1) + dx_{t1}\cos(\omega_1 t_1)]}{\sin(\omega_1 t_1)} - \frac{dx_{t1}}{\sin(\omega_1 t_1)} && (2.30)\\
&= (x_{t1} - x_0)\cos(\omega_1 t_1) + \frac{dx_{t1}\cos^2(\omega_1 t_1)}{\sin(\omega_1 t_1)} - \frac{dx_{t1}}{\sin(\omega_1 t_1)} && (2.31)\\
&= (x_{t1} - x_0)\cos(\omega_1 t_1) + \frac{dx_{t1}[\cos^2(\omega_1 t_1) - 1]}{\sin(\omega_1 t_1)} && (2.32)
\end{aligned}$$

Fig. 2.3 shows $x_1(t)$ drawn in blue.

Displacement of the upper part of the vocal folds. The displacement $x_2(t)$ of M_2 is given in [Stev 98] as

$$x_2(t) = x_{20}(1 - \cos(\omega_2(t - t_1))) \tag{2.33}$$

where ω_2 is the natural frequency of mass M_2: $\omega_2 = 1/\sqrt{C_c M_2}$. The amplitude x_{20} of $x_2(t)$ is determined by the difference between the peak displacement of M_1 and x_{10} [Stev 98]. The point $t_{\max}$ for which $x_1(t)$ reaches its maximum is

$$t_{\max} = \left[\arctan\left(\frac{k_2}{k_1}\right) + \pi\right] \cdot \frac{1}{\omega_1} \tag{2.34}$$

resulting in:

$$x_{20} = k_2\sin(\omega_1 t_{\max}) + k_1\cos(\omega_1 t_{\max}) + x_0 + C_1 P_s d_1 - x_{10} \tag{2.35}$$

where the term containing P_s can be ignored as $P_s = 0$ for $t > t_1$. The parameters k_1, k_2 are given by Eq. 2.32 and Eq. 2.28. Fig. 2.3 shows $x_2(t)$.

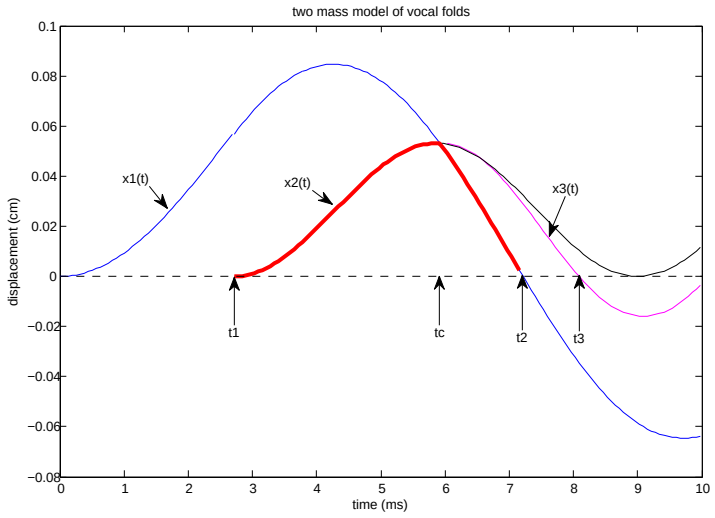

Figure 2.3: Displacement of the two masses over time.

Glottal opening. As the air has to pass both the lower and the upper part of the vocal folds, the glottal distance $d(t)$ is proportional to the minimum opening of both parts. It can easily be seen from Fig. 2.3 that the glottis is opened only from t_1 to t_2, the point where $x_1(t)$ becomes zero again. As the vocal fold model is symmetric, we have to multiply the displacements by two in order to get the glottal width $d(t)$:

$$d(t) = 2 \cdot \min(x_1(t), x_2(t)) \tag{2.36}$$

The minimum of $x_1(t)$ and $x_2(t)$, i.e. the part of the glottal cycle where the glottis is open, is shown in red in Fig. 2.3.

The acoustic mass of the air in the glottis, in the trachea and in the vocal tract below and above the glottis results in a skewing of the airflow waveform. The higher the constriction of the vocal tract, the larger the acoustic mass and the larger the skewness of the waveform. The acoustic mass is not taken into account by the glottal airflow computation described in the next section. In order to provide more possible variability to the shape of the excitation pulse, we added a heuristic skewness transform to the displacement signal. The skewness transform is applied prior to the calculation of the airflow waveform U_g.

$$d(t) = d(t) \cdot \phi \cdot t + d(t) \tag{2.37}$$

The greater the parameter ϕ, the higher the skewness of the transformed displacement function. Thus, one may think of it as an approximate representation of the constriction of the vocal tract.

The area of the glottal opening $A_g(t)$ can be easily approximated from the glottal width $d(t)$ and the length of the glottis l assuming a rectangular shape:

$$A_g(t) = l \cdot d(t) \tag{2.38}$$

Length of the closed phase. Once $x_1(t)$ has reached zero at t_2, the glottis remains closed until the whole glottal opening-closing cycle restarts at t_0. Thus, the glottis is closed not

only at the beginning of the cycle from t_0 to t_1, but also at its end in the interval from t_2 to t_3, where t_3 denotes the end of the cycle. This is illustrated in Fig. 2.3. In order to be able to compute the overall length of the cycle we need to approximate t_3, the point when the upper part of the vocal folds closes. The definition of $x_2(t)$ in Eq. 2.33 does not take into account the influence of the inward movement of the lower part of the vocal folds which gets more rapid towards the end. The larger slope of $x_1(t)$ is most prominent after the intersection of $x_1(t)$ and $x_2(t)$ at t_c. This means, $x_2(t)$ should have a much steeper decent after t_c and $x_2(t)$ as defined in Eq. 2.33 cannot be used to compute t_3. Therefore we approximate $x_2(t)$ for $t > t_c$ by a function $x_3(t)$ which has a steeper slope than $x_2(t)$. In order to ensure a smooth transition from $x_2(t)$ to $x_3(t)$ at t_c we define $x_3(t)$ by the following equations:

$$j_2 = (x_2(t_c) - \zeta \cdot x_{20}) \cdot \sin(\omega_2(t_c - t_1)) + x_2'(t_c) \cdot \cos(\omega_2(t_c - t_1)) \quad (2.39)$$

$$j_1 = \frac{\cos(\omega_2(t_c - t_1)) \cdot ((x_2(t_c) - \zeta \cdot x_{20}) \cdot \sin(\omega_2(t_c - t_1)) + x_2'(t_c) \cdot \cos(\omega_2(t_c - t_1))) - x_2'(t_c)}{\sin(\omega_2(t_c - t_1))} \quad (2.40)$$

$$x_3(t) = j_2 \cdot \sin(\omega_2(t - t_1)) + j_1 \cdot \cos(\omega_2(t - t_1)) + \zeta \cdot x_{20} \quad (2.41)$$

The parameter ζ controls the steepness of $x_3(t)$. As $x_3(t)$ replaces $x_2(t)$ *after* the crossing at t_c, it has no influence on $d(t)$; it is solely used for the computation of t_3 which is the point at which $x_3(t)$ is zero.

2.2.2 Glottal Airflow

The next step is to compute the volume velocity $U_g(t)$ of the glottal airflow from the glottal opening. The relationship between the volume velocity $U_g(t)$ and the transglottal pressure drop ΔP_g can be found in [Stev 98]:

$$\Delta P_g = \frac{12\mu h}{ld^3(t)} U_g(t) + \kappa \frac{\rho}{2(ld(t))^2} U_g^2(t) \quad (2.42)$$

The parameter μ represents viscosity, h the thickness of the glottal slit, κ depends on the shape of the glottal slit, and ρ is the density of air. The glottal width $d(t)$ is given by Eq. 2.36 and Eq. 2.37 in the previous section. Here we assume that ΔP_g is approximately constant over time. Solving for $U_g(t)$ using the quadratic equation leads to

$$U_g(t) = \frac{-\frac{12\mu h}{ld^3(t)} \pm \sqrt{(\frac{12\mu h}{ld^3(t)})^2 - 4\kappa \frac{\rho}{2(ld(t))^2} \Delta P_g}}{2\kappa \frac{\rho}{2(ld(t))^2}} \quad (2.43)$$

The derivative of $U_g(t)$ yields the excitation pulse. In our experiments we employed a discrete time approximation of the derivative. Fig. 2.4 illustrates the three steps of the computation of the excitation function: the area of the glottal opening $A_g(t)$ (shown on the left) is computed, then the volume velocity $U_g(t)$ (middle) is derived from the glottal opening, and third, applying the time derivative leads to the excitation pulse shown on the right. The above analysis results in an excitation pulse for a single pitch. Of course, speech contains varying pitch frequencies. The model parameters must be adapted to account for various pitch values. Variation of the parameters can produce different shapes of the excitation signal as well. For instance, Fig. 2.5 shows three excitation pulses for three different parameter

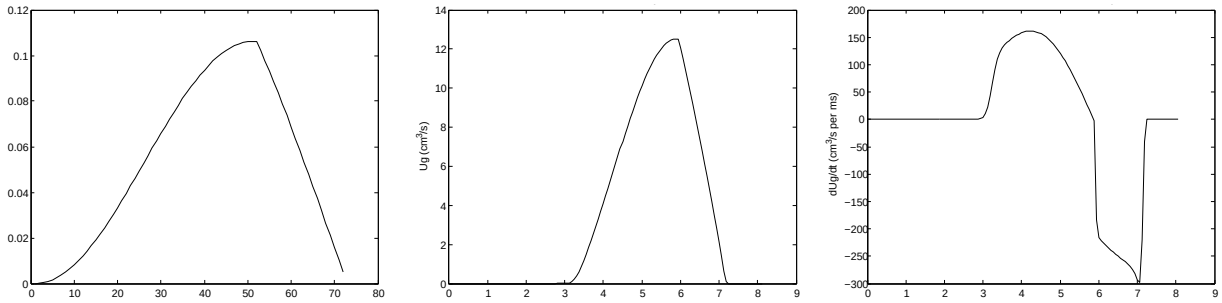

Figure 2.4: Computation of the excitation function in three steps. $A_g(t)$ (left) $U_g(t)$(center) excitation pulse (right)

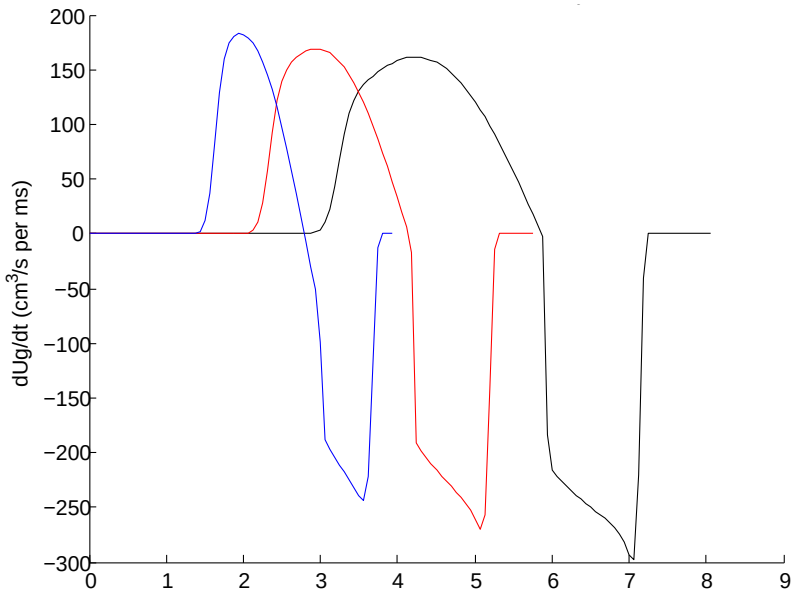

Figure 2.5: Excitation function for three different parameter settings.

settings. It is obvious that not only the pitch frequency changes but also the shape of the excitation signal. In the next section we investigate fitting the model parameters to match the specific characteristics of a speaker's voice at a certain age.

2.3 Model Optimization

Our hypothesis is that glottis model parameters contain information about speaker age. To test this hypothesis, we find the optimal model parameters that fit the speech data and observe how they change with age. Importantly, we are interested in model parameters that are independent of pitch. However, the generated pitch is determined by the glottis model parameters. Therefore, in our analysis we observe the model parameters across age *only for a single pitch.* (In practice, we use a small pitch range of 10 Hz around the pitch of interest.) In order to accomplish this, we optimize the model parameters for every 25 ms speech frame. Then we are able to sort and analyze the model parameters according to the pitch for each

25 ms frame.

Figure 2.6 depicts a block diagram of the optimization loop. To begin, a set of initial

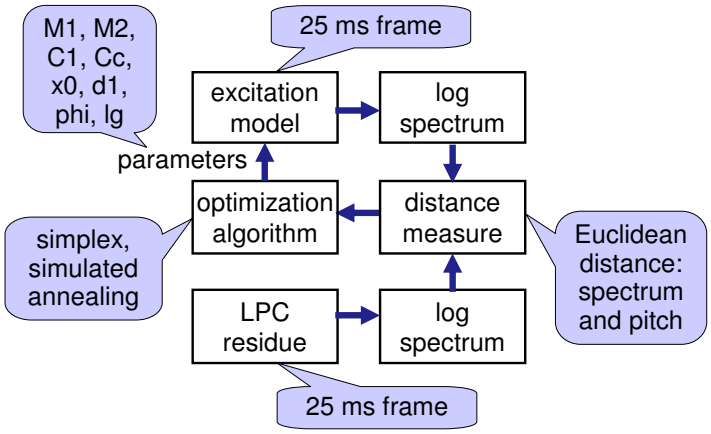

Figure 2.6: Optimization Loop

parameters (M_1, M_2, C_1, C_c, x_0, d_1, ϕ, l) is input into the glottis excitation model. Given the parameters, the model generates an excitation signal for a 25 ms speech frame. At the same time, we calculate the LPC residue of the original speech signal. Then we apply a log spectrum transform to both of these excitation signals. The similarity of the generated excitation signal is compared to the original signal using two Euclidean distances. First, we compare the distance between the log spectrum of the two signals. Second, we compare the distance between the generated and the original pitch for the frame. The combined distance measure is passed to the optimization algorithm, which modifies the parameter set, passing the new parameter set to the excitation model. Thus, an optimization loop is formed, modifying the parameters, generating a new candidate excitation signal, and testing it against the original signal. We used two different optimization algorithms (simplex algorithm and simulated annealing) in order to find the set of parameters that minimized the distance between the generated and original signals for every 25 ms speech frame.

The optimization is formulated as:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmin}}[D(s_m(\theta), s_{org})] \\ \theta &= \{M_1, M_2, C_1, C_c, x_0, d_1, \phi, l\} \end{aligned} \tag{2.44}$$

where $D(s_m(\theta), s_{\text{org}})$ is the combined distance between the model excitation signal s_m and the original excitation signal s_{org}. The combined distance measure combines distances between both the respective log spectra and the respective pitches p_m, p_{org}, and is defined as:

$$D(s_m(\theta), s_{\text{org}}) = D(\text{logspec}(s_m(\theta)), \text{logspec}(s_{\text{org}})) + \lambda \cdot D(p_m, p_{\text{org}}) \tag{2.45}$$

where $D(\cdot, \cdot)$ is the Euclidean distance between two vectors and the constant λ scales the influence of the pitch distance.

An example of optimizing the excitation signal for a 25 ms frame is shown in Figure 2.7. The left figure shows the log spectrum of the original signal (blue) and the modeled excitation signal (red) using the default start parameters. Following optimization, the log

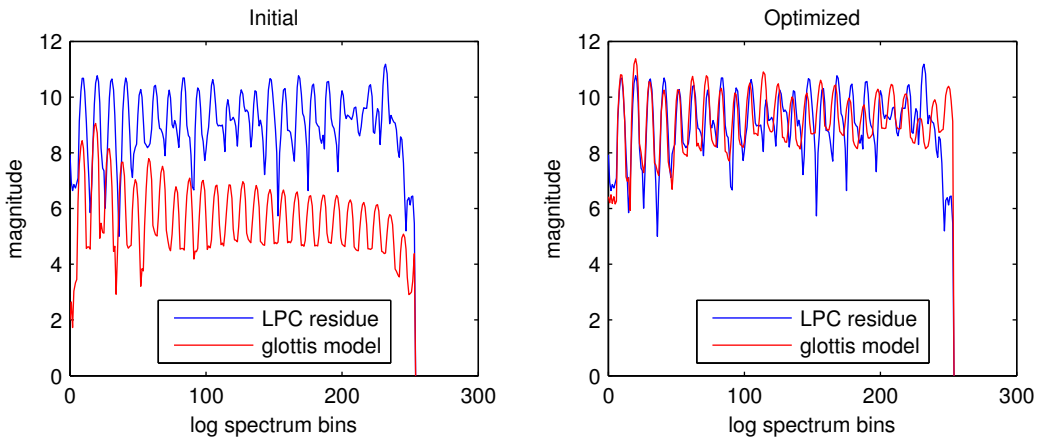

Figure 2.7: Excitation signal optimization – log spectrum

spectra are shown in the right figure. The first peak is the pitch, and it matches almost exactly. Not all of the higher order harmonics match closely, however, the overall shape of the two log spectra are well matched. (Note that the log spectrum transform contains a pre-emphasis filter, giving the spectrum its generally uniform magnitude across frequencies.)

Our approach is to analyze the excitation model parameters for a single pitch, as they vary across age. We found the most frequent pitch for the speaker over all ages, and used that pitch for all subsequent analyzes. Figure 2.8 shows histograms of pitch values for the Queen and Cooke over all ages. The histogram maximum is 205-215 Hz for the Queen and 85-95 Hz for Cooke. Note that the optimization is only performed on voiced speech segments.

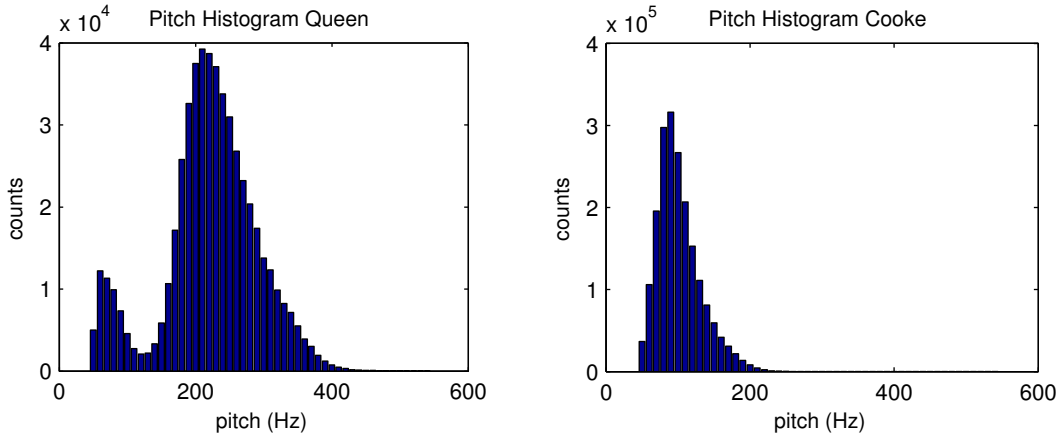

Figure 2.8: Pitch Histograms

2.3.1 Simplex Algorithm

We used two different parametric search algorithms in the optimization loop to find the optimal model parameters. The first parametric search algorithm is the Nelder-Mead simplex algorithm [Neld 65, Olss 75]. The method is based on a simplex (or flexible polyhedron) that traverses the multidimensional design space. For each iteration, a new trial point is generated

outside of the simplex. The trial point is determined taking a step in the opposite direction of the largest valued point on the simplex (reflection computation). Based on the value (objective function distance) of the trial point in comparison to the current simplex values, the simplex either extends, contracts, or shrinks. In this manner, the simplex descends the gradient of the loss function, expanding, contracting, and shrinking until it converges on a minima. One caveat is that, since the algorithm descends the gradient, it is only capable of finding local minima.

2.3.2 Simulated Annealing Algorithm

The second parametric search algorithm used in the optimization loop is the simulated annealing algorithm [Kirk 83]. This stochastic search algorithm randomly selects a trial point and calculates the objective function distance at that point. If the distance is smaller than the current distance, the point is accepted and a new iteration cycle begins. If the distance is larger, however, the trial point is accepted with some probability $P \propto \exp(-d_{trial}/k_B T)$, where k_B is a constant and T is the "temperature" of the system. In this manner, it is possible for the algorithm to climb as well as descend the objective function gradient. This leads to a higher likelihood that the search will find the global optimum, however this is not guaranteed. The temperature of the system refers to the analogy with physical annealing, where an amorphous solid is heated until it melts and then is slowly cooled. After cooling is finished the material solidifies with a crystalline structure. In the case of the simulated annealing algorithm, at high temperatures the algorithm is more likely to climb the gradient, while at low temperatures it will not. Hence, the algorithm proceeds according to a cooling schedule (exponentially decreasing temperature in our implementation.) In addition, we also modify the step size based on the temperature of the system. At high temperatures, the random trial points are selected from a wide range. As the temperature cools, the range is reduced so that the algorithm searches its local neighborhood. In this way, the algorithm begins by making coarse steps over the search space and ends by making small steps, converging to a higher resolution final result.

2.4 Results

We performed three experiments on three different corpora, the Queen of England, Alistair Cooke, and the University of Florida Vocal Aging Database (UF-VAD). The Queen corpus is from 30 Christmas speeches addressed to the people of England over the years 1952 to 2002. The Queen's age ranges from 26 to 76 in the recordings. The total length of the corpus is approximately 2.5 hours of speech. The Cooke corpus consists of 30 "Letter from America" radio broadcasts by Alistair Cooke. Recorded over the years 1947 to 2003, the broadcasts cover Cooke from age 38 to 95. The total length of the corpus is approximately 10.6 hours of speech. The Queen and Cooke corpora are longitudinal data, covering a single speaker over many years. The UF-VAD corpus contains cross-sectional data, covering many speakers of many ages, all recorded in the same time period with the same recording equipment. There are 150 speakers ranging from ages 18 to 92, all recorded from 2003-2007.

2.4.1 Queen

We analyzed the excitation model parameters for the Queen across all ages in the pitch range 215-235 Hz. The correlation between each parameter and the age of the Queen is summarized in Table 2.2. The highest correlations are for the parameters M_2, x_0, and

Table 2.2: Queen Results

	M_1	M_2	C_1	C_c	d_1	x_0	ϕ	l
simplex	0.29	0.54	0.36	0.49	-0.33	0.64	0.55	-0.59
simulated annealing	0.53	0.49	0.47	0.00	-0.20	0.10	0.36	0.34

ϕ, optimized by the simplex algorithm, and M_1, M_2, and C_1 for the simulated annealing algorithm. The optimal M_2 parameter is plotted versus age for both algorithms in Figure 2.9. The red line is the least squares regression fit line to the data. From observing the

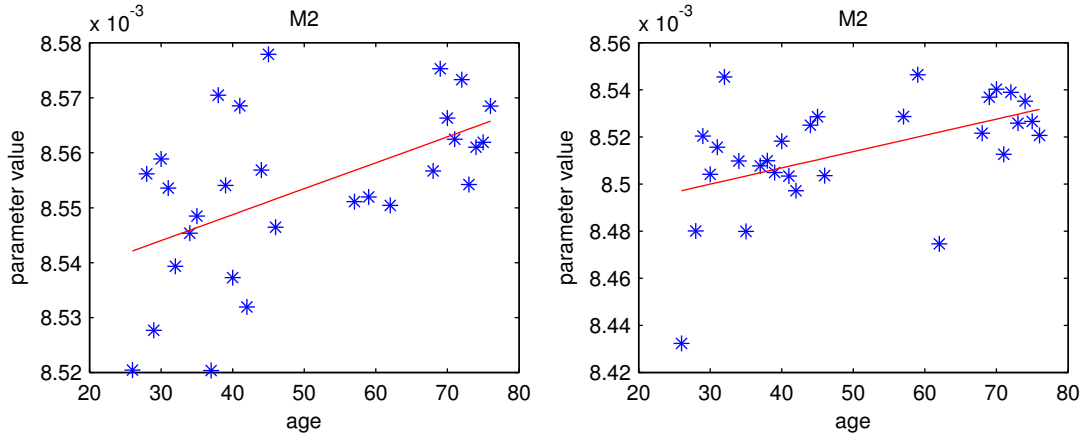

Figure 2.9: Queen Results: Parameter M_2 vs. Age. Simplex Algorithm (left) Simulated Annealing Algorithm (right)

plots, there is a generally increasing trend in M_2 as age increases. However, there is a large variance in the M_2 parameter, resulting in the moderate correlation value.

2.4.2 Cooke

The same analysis was performed on the Cooke dataset using a pitch range of 75-95 Hz. Table 2.3 summarizes the correlation between the excitation model parameters and age. In this case, we split the correlation analysis into two regions, one when Cooke was younger than 80 years old and the other region when he was 80 years or older. This resulted in very strong correlations for the optimal M_1, M_2, and C_1 parameters for both age regions and both optimizing algorithms. The optimal M_1 parameter versus age is plotted in Figure 2.10 for both optimizing algorithms. The least squares regression lines are shown in red. In this case, the parameter variance is low and the correlation of the parameter with age is very clearly

Table 2.3: Cooke results

	age	M_1	M_2	C_1	C_c	d_1	x_0	ϕ	l
simplex	< 80	0.90	0.44	0.92	0.62	-0.01	0.25	0.38	-0.73
	≥ 80	-0.79	-0.77	-0.88	-0.57	-0.61	-0.48	-0.33	0.47
simulated annealing	<80	0.92	0.81	0.75	0.82	-0.33	-0.03	0.24	0.06
	≥80	-0.89	-0.88	-0.90	-0.83	0.03	-0.58	-0.65	0.33

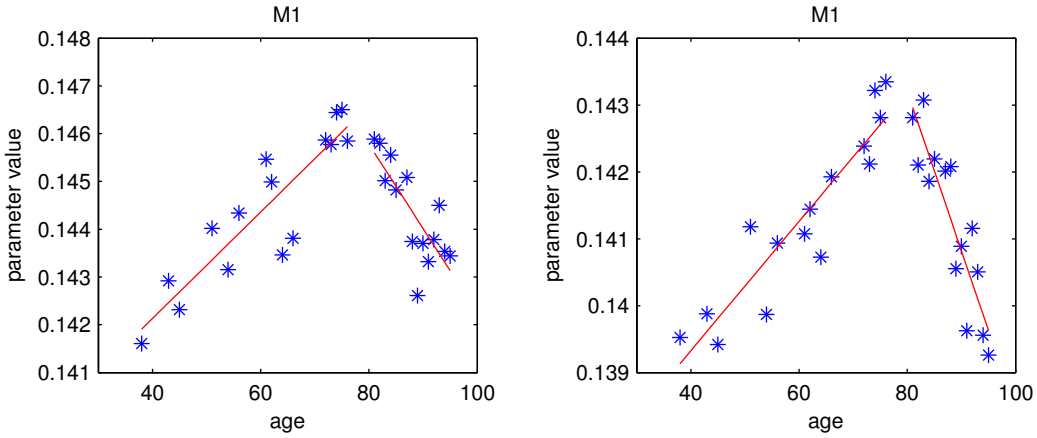

Figure 2.10: Cooke Results. Simplex Algorithm (left) Simulated Annealing Algorithm (right)

seen in both regions. One important difference between the Queen and Cooke experiments is the amount of data. The Cooke corpus is approximately four times as long as the Queen corpus. This probably plays a significant role in the larger amount of noise variance in the Queen optimal parameter analysis. This hypothesis could be tested by reducing the length of the Cooke corpus and observing the effect on variance.

2.4.3 UF-VAD

The goal of the third experiment was to determine if there is *speaker independent* age information contained in the glottis model parameters (in addition to pitch and formants). The approach was to create an age predictor using support vector regression (SVR) with a linear kernel. The age predictor was trained and tested on the multiple speaker UF-VAD corpus. To form features for the SVR, we synthesized the glottis excitation signal from the optimal (simplex) model parameters for each speaker. Then we transformed the synthesized excitation signal into MFCCs and used the MFCCs as features to train and test the SVR.

The results are summarized in Table 2.2. As a baseline, SVR was performed on f0 and the first three formants of the speech signals. This had a low correlation for male speakers, but relatively high correlation for female speakers. Using 36 MFCCs from the re-synthesized glottis excitation signal had much better correlation with age than the baseline system for males, and comparable correlation for females. For both males and females, combining the f0, formant, and glottis model features, resulted in the highest correlation with age.

Table 2.4: SVR results

	f0 + 3 formants	glottis (MFCCs)	glottis (MFCCs)	glottis + f0 + 3 formants
Dimension	4	24	36	24
Male: Correlation	0.32	0.36	0.58	0.66
Male: Mean Abs Err	18.95	21.77	14.4	17.05
Female: Correlation	0.77	0.70	0.74	0.80
Female: Mean Abs Err	13.08	15.65	12.34	14.25

2.5 Conclusion

In summary, we have implemented the Stevens model of the glottis to synthesize glottal excitation signals based on eight parameters. Using short time frames (25ms), we minimized the distance between the synthesized excitation signal and the LPC residue to find optimal model parameters. These per frame parameters were averaged for a single pitch range and a single age. Then we analyzed the correlation of the model parameters with age. Two experiments on longitudinal data show correlation between the model parameters and age. A third experiment applied support vector regression to predict speaker age, using the re-synthesized glottal excitation signal as features. All three of these experiments provide evidence that (1) age information is contained in the glottal excitation signal and (2) that information may be extracted from a reduced physical model of the glottis with few parameters.

Future research will focus on: refining the models, parameters and experiments, (for example, include a breathiness parameter) and second, applying the approach to more data.

Chapter 3

Error Analysis of Formant Tracking Algorithms

3.1 Motivation and Goal

Several subgroups (inversion, feature extraction for age classification, indirectly also speech recognition) used formant frequency values as input to their analyzes. Most of all, the acoustic-to-articulatory inversion relied heavily on the formant frequencies in order to determine possible vocal tract shapes for re-synthesis. We wanted to find *robust* methods for vocal aging related issues so automatically extracted formant values using formant tracking algorithms were considered as a prerequisite for fast and robust data processing.

We conducted an error analysis of the formant tracking algorithms of several softwares (WinSnoori, Wavesurfer) to find out whether the automatically extracted formant frequencies were influenced by age. Automatic tracking algorithms always make systematic errors since formant extraction is not a trivial task. The question was whether these systematic errors changed due to the age of the voice that was being analyzed. Using the *Queen Christmas Speeches* corpus, we manually annotated a subset of it and compared it to the automatically extracted formant frequencies of two softwares. This section first describes the procedure to build the reference corpus within the manual annotation task, and the settings of how the automatically tracked formants were obtained. After that, the error analysis itself is being described.

3.2 Manual Annotation

3.2.1 Selection of Material for the Reference Corpus

We chose a subset of wave file "chunks" of the *Queen* corpus. With a real-time factor of 1:120 (30 secs annotation take about 60 mins), we limited ourselves to a subset of the original corpus (approx. 11 mins or 13 %) that was distributed over the decades. We also took not only audio chunks from the beginning of a given speech but also from the third minute of speech to account for possible changes in speaking style within a speech. The speech material selected for the manual annotation corpus can be seen in table 3.1.

Year	1952	1957	1966	1972	1983	1994	2002	All years
Chunk	1st, 3rd	1st, 3rd	3rd	1st, 3rd	1st, 3rd	3rd	1st, 3rd	
Mins	1.55	2.28	0.85	1.75	2.35	1.02	1.85	11.65

Table 3.1: Manually annotated speech material

3.2.2 Annotation Procedure

Our baseline for annotation was *Wavesurfer* (`http://www.speech.kth.se/wavesurfer`, version 1.85), calculating the first 4 formant frequencies by using the standard settings. Two trained annotators manually corrected the proposed formant trajectories of the software by visual inspection of the spectrogram.

This meant a bias in the error analysis – but we were not so much interested in absolute error but rather in relative errors over age, see below.

The annotators repeatedly discussed their decisions to reach an agreement on the corrections and to increase the quality of the annotation.

The visual inspection task of the spectrogram was based on the following guidelines and phonetic background knowledge:

- Vowel formant charts for general overview of absolute positions of formants in the frequency domain
- Listening (for vowel reductions etc.)
- Time alignment of the segments for faster orientation in the speech signal
- Consonant-vowel transition charts to determine formant transitions, where hard to see
- Pitch contour display using Wavesurfer's default settings alongside with the spectrogram to identify speech parts classified as "voiced" by Wavesurfer

The decision procedure of *what* to correct and *how* can be summarized as follows:

- Only look at voiced segments that are non-nasalized and no consonants (i.e. vowels), since only for vowels the values make sense
- Ignore *fine* errors due to f0/F1 interaction (due to lower bound of the resolution of the computer screen and the mouse pointer)
- Focus on correcting *coarse* errors (when formant tracker picked the "wrong" formant) (see Figure 3.1)

The task then was to re-draw and smooth the formant contour correctly as derived from the visual inspection of the spectrogram (dark bars indicating the high energy bands of the resonance frequencies, i.e. formants).

Problems mainly occurred with weak signals especially in older recordings which made it hard to decide on the actual formant contour; in this case the lines were not corrected since it would not necessarily have meant any "improvement".

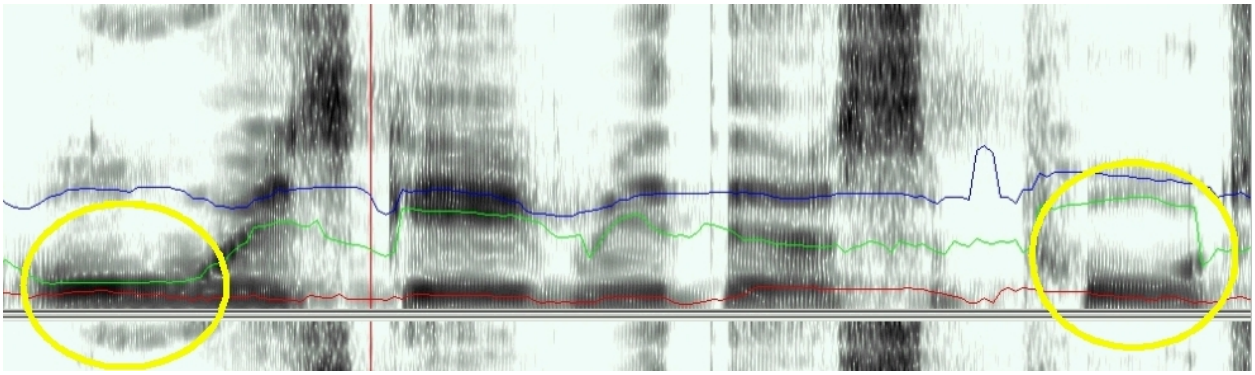

Figure 3.1: The same vowel was tracked correctly (left) and incorrectly (right). F2 should be rather low for /u:/. In the right example, a typical coarse error occurred: The tracking algorithm chose the F3 energy band as being "F2". Manual correction would redraw the green line in a lower position.

3.2.3 Automatic Extraction

For the error analysis we used the same subset of the *Queen* corpus as prepared in the manual annotation task. The softwares used for automatic extraction were

- Wavesurfer (http://www.speech.kth.se/wavesurfer)
- WinSnoori (http://www.loria.fr/∼laprie/WinSnoori/index.html)

Frame interval was set to 10 ms in each analysis software to keep the values comparable (i.e. at the same time stamps in the signal). Although we only considered the first 3 formants in this analysis, the number of formants automatically tracked was set to 4. Setting it to 3 would have substantially deteriorated the quality of the formant tracking results.

3.3 Error Analysis of Formant Values

We compared the manually annotated formant frequencies against automatically extracted values. The data that was taken into account only came from the "voiced" segments as determined by the pitch tracker of Wavesurfer.

For each frame of the speech signal, we calculated the absolute difference in Hertz between the manually annotated and automatically extracted formant frequencies. We calculated Root Mean Square (RMS) error, standard deviation, and median. We also calculated the percentage of the speech material where the formant values matched or were very close to each other (in the range of ± 32Hz). This tells us how well the automatic formant extraction algorithms are doing. As can be seen in Table 3.2, the correct or closely matching values for Wavesurfer are a lot higher than for WinSnoori. This is due to the fact that the tracking output of WaveSurfer was used as starting point for the manual annotation.

We then plotted histograms for the difference values. We tried different bin sizes and different ranges of the lower and upper limit of the histograms. We observed no systematic error in automatically tracked formant values over age in any case. An example can be seen

Table 3.2: *Percentages of exactly and closely (i.e. ± 32 Hz) matching formant frequencies after manual correction (averages and standard deviations over 17 files each)*

	. . . vs. WaveSurfer			**. . . vs. WinSnoori**			
	F1	F2	F3	F1	F2	F3	
Manually corrected . . .	88,14	86,24	86,76	24,14	14,94	16,67	(Ave)
	4,03	6,51	7,84	6,2	1,32	4,19	(SD)

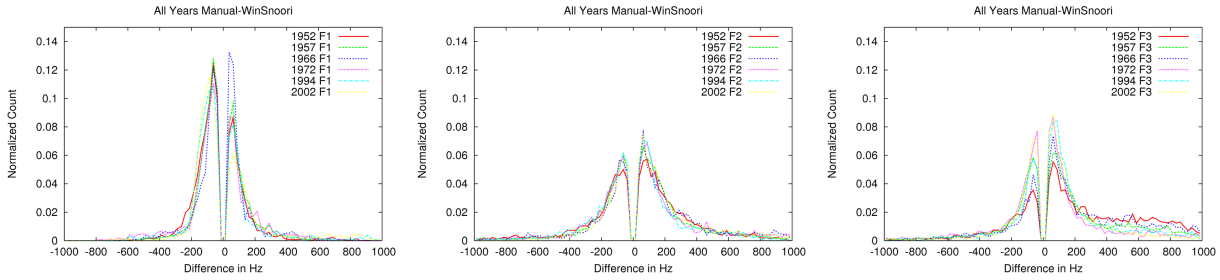

Figure 3.2: Difference value histograms of F1 (left) to F3 (right) of *WinSnoori*. Bin size = 80; range = ±1000 Hz.

in Figure 3.2. We do not plot the (very high) peak values around the matching values (cf. Table 3.2) but instead enable a focus on the mismatches to either side.

3.3.1 Description

On the x-axis, we see the difference of the manually corrected minus the automatically extracted formant frequencies. E.g. when the manual value suggests 600 Hz for F1, and an automatically extracted value shows 400 Hz, the difference is +200 Hz, being shown on the *right* side of the peak. This means that the automatic tracking algorithm underestimated the formant frequency for this example. The underestimation *can* be a systematic error but since it is the same over all years, it is not seen as an influence of age.

3.3.2 Interpretation

For F1 and F2, the differences in the formant frequencies are quite homogeneous over the years. We see a slight increase in "noise" in the case of F3: The recordings of the young Queen show the largest differences compared to the manually corrected values but the decrease of the differences was not strictly associated with age. Thus, the overall conclusion and working assumption for our project was that we could indeed regard the automatic formant extraction as a robust method to provide input values to our further analysis.

3.4 Correlation Results

We know that the manually annotated data covered only 13 % of the whole *Queen* corpus as the annotation is a time consuming task. However, we believe that 11 minutes of data

Table 3.3: *Pearson and Spearman correlations with age*

	Pearson			Spearman		
	F1	F2	F3	F1	F2	F3
Manual	-0.95	-0.82	-0.12	-0.96	-0.85	-0.11
Wavesurfer	-0.92	-0.70	0.40	-0.89	-0.60	0.42
WinSnoori	-0.96	-0.63	0.32	-0.96	-0.57	0.39

is not too little for the analysis. Moreover, we chose well distributed data over age. Thus, although the amount of manually annotated data was comparatively small, we calculated the correlation of hand corrected formant frequencies with age. Table 3.3 shows the correlation values for comparison. In the manual case, F1 shows a strong negative correlation with age, as does F2. F3 seems to be uncorrelated in our data set. In the automatically tracked cases, we observe that the strong correlation for F1 with age is confirmed. F2 is not as strongly correlated. The correlation result of F3 in the manual case is not confirmed in the automatic case where we observe a slightly positive correlation.

3.5 Limitations

Since formants are normally used to describe vowels, the manual correction only made sense for vowels; nevertheless, voiced consonants were also part of the error analysis. This was due to the fact that the pitch tracking algorithm detected pitch in voiced consonants as well and we therefore included *all* voiced segments into the error analysis. The annotation took the values of Wavesurfer as a starting point and therefore increased the bias towards Wavesurfer: Not correcting "formant" values of voiced consonantal segments means that, in the error analysis, they count as perfectly tracked formants in the case of Wavesurfer, but not in the case of WinSnoori. However, we did not evaluate the tracking algorithms against each other but were only interested in *relative* changes over the years.

Chapter 4

Vocal Tract Inversion

4.1 Maeda's Articulatory Model

One of the most important aspects in acoustic-to-articulatory inversion is using an articulatory model that is capable of describing all useful vocal tract configurations, and hopefully only useful configurations.

Many articulatory models have been proposed in the literature, from the most simple (Fant's articulatory model [Fant 60], is described by only 3 control parameters), to the most complex (Wilhelms-Tricarico's model is a 3D-biodynamical model of the tongue using finite elements, which is controlled by hundreds of parameters).

One of the most famous is Maeda's articulatory model [Maed 79, Maed 90]. He describes a full vocal tract using three independent models for the lips, the tongue and the larynx. These three articulators indeed may be considered to have independent influences, although all three are influenced by the position of the lower jaw. The factorial analysis used by Maeda to derive the model had to be powerful enough to take this particularity into account and subtract the influence of the lower jaw to study the other articulators.

The position of the lower jaw can easily be derived on the X-rays images by measuring the distance between the lower and upper incisive. Principal Component Analysis is not adequate in this case; Maeda [Maed 79] thus used a different method, known as arbitrary orthogonal component analysis [Over 62], to subtract the influence of the jaw position. Each region of the vocal tract (lips, tongue, larynx) can then be studied independently.

For each region, control parameters are derived by applying a principal component analysis on the data decorrelated of the influence of the jaw, and keeping enough components to explain the majority of the variance. The number of necessary parameters vary in each region of the vocal tract; for the larynx, one parameter is enough; for the lips region, three variables are analyzed: vertical opening of the lips, horizontal opening of the lips (or stretching), and protrusion. Two intrinsic parameters were kept to describe these variables: vertical opening and protrusion; the horizontal opening is deduced from these two parameters. For the tongue, three additional parameters are necessary to explain 96% of the variance on the X-ray images; in total, the model is controlled by 7 parameters (cf. fig. 4.1).

Furthermore, the model can be adapted to different speakers: two scaling parameters for the oral and pharyngeal tracts allows the adaptation of the shape of the tract to a new

speaker. These two scaling parameters have homogeneous influence on the dimensions of the two tracts but with an adequate procedure, it is possible to have a model capable of mimicking the acoustic production of an arbitrary speaker. Galvan-Rodrigez established a semi-automatic method to adapt the model from the formants frequencies of specific vowels for a given speaker [Galv 97a]. This method however make the hypothesis that a specific vowel is produced by a unique articulatory configuration independently of the speaker.

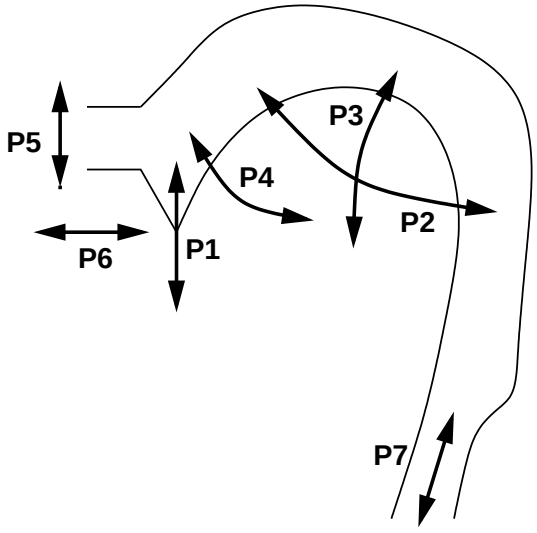

Figure 4.1: The seven parameters of Maeda's articulatory model: the jaw (or jw) P1, the vertical opening of the lips (lh) P5, the lip protrusion (lp) P6, the tongue body position (tb) P2, the tongue shape (ts) P3, a parameter controlling the tongue tip (tt) P4, and finally the larynx height (lx) P7.

4.2 Inversion by Variational Calculus

For this workshop a novel method of inversion based on variational calculus is presented to solve the dynamic acoustic-to-articulatory mapping without the need for an initial solution. It finds its root in a work proposed by Laprie and Mathieu [Lapr 98], but extends so it does not require an initial solution.

Laprie and Mathieu, similarly to e.g. Schoentgen and Sorokin, minimize a cost function which is a combination of pseudo-potential and pseudo-kinetic energy terms. A first term of the cost function is $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$, which expresses the changing rate of articulatory parameters, and in order to penalize large articulatory efforts and prevent the vocal tract from reaching positions too far from equilibrium, a potential energy term $\sum_{i=1}^{7} k_i \alpha_i^2(t)$ is added.

The cost function to be minimized has the following form:

$$I = \int_{t_i}^{t_f} \sum_{j=1}^{3} (f_j(t) - F_j(\alpha(t))^2 + \lambda \sum_{i=1}^{7} m_i \alpha_i'^2(t) + \beta \sum_{i=1}^{7} k_i \alpha_i^2(t) \, \mathrm{d}t \tag{4.1}$$

The main difference of the method we used in the workshop is the way the acoustic criterion is handled. In the work of Mathieu, the acoustic criterion was of the form : $\sum_j (f_j(t) - F_j(\alpha(t)))^2$. Unfortunately, this criterion is subject to numerical instability. Sorokin [Soro 00] proposes to use a slightly different criterion: $\max_j |1 - F_j(\alpha(t))/f_j(t)|$ which is not very practical because this function has singular derivatives. A compromise form is the following: $\sum_j (1 - F_j(\alpha(t))/f_j(t))^2$ (which is the same as Sorokin, replacing the infinite norm by the Euclidean norm). This form is much more numerically stable, but it is still only valid in a small vicinity of a correct solution, so it can only be used in conjunction with an initial solution. An acoustic criterion which always converge towards a correct solution is thus needed.

The acoustic criterion has thus been replaced by a more efficient one. An explicit correction vector is computed, based on the use of the pseudo-inverse computed similarly to Schoentgen [Scho 97] through the use of Singular value Decomposition, allows to explicitly compute a correction vector that is more likely to converge towards a local minimum of the acoustic error.

4.3 Discussion

This method has proven its effectiveness for acoustic-to-articulatory inversion using the first formants frequencies has input. It is fairly simple, and is real-time when combined with a high quality codebook synthesizer such as the one presented in [Pota 07].

It however lacks robustness in the case of "mistakes" in the input acoustic vector: "impossible" acoustic vectors can sometime produce a large error in the articulatory trajectory, and should preferably be eliminated; this problem is even more frequent when using more complex input acoustic vectors, such as LPC coefficients.

4.4 Experiments and Results

Inversion experiments were conducted on several speech corpus: without any specific adaptation of the model, articulatory parameters were obtained to describe the most likely vocal tract shapes that produced the original speech signal. The inability to conduct a model adaptation to the speakers was due to the large number of speakers to process, the time constraints, and the fact that the current method is adapted to the French language. This implies that some of our assumptions are wrong, and that the absolute values found for the parameters are probably meaningless. We however expect their evolution over time to be fairly meaningful.

After obtaining the trajectories of the articulatory parameters P1...P7 over time, some basic statistics were performed, and the evolution of the average value and the standard deviation for all seven parameters were observed. Some of the parameters demonstrated some interesting trends, but others (e.g. P7, the larynx height parameter) showed unexpectedly no evolution.

4.4.1 Queen

The main subject of our study was H.M. the Queen Elizabeth II. The corpus was the Christmas broadcast speech that she pronounces every year, over a time period of 50 years. This is a good corpus to study the evolution of the articulatory parameters with age, since all the recordings are from the same speaker. The recording conditions were however quite different from one year to another.

Fig. 4.2 displays the evolution of parameter P1 (the jaw parameter) over age. Lower values mean a lowering of the lower jaw, i.e. a wider opening. We observe that over age, the average of this parameter is increasing, which means that on average, the mouth is less open. A more careful observation of the trajectories shows that the maximum values for this parameter is stable with age, but the smallest value is increasing with age. Fig. 4.2 show that the standard deviation for this parameter is also decreasing quite a lot with age, which indicate that the amplitudes of the movements are lower, which seem to be an indication of a change due to aging.

4.4.2 Adaptation of the articulatory model by Galvan's method

This method, proposed by Galvan and Naito, use two scale parameters to adapt the articulatory model to any speaker. The two scale parameters - normalization factors for the lengths of respectively the oral and pharyngeal tubes - are derived using the method described in [Galv 97b]. A simplified version of this method is described in [Nait 99].

The idea of this method is fairly simple, although the theoretical background it relies on is debatable. To simplify, it is assumed that each phoneme is pronounced with the same articulatory configuration by any speaker; what changes the acoustic is merely the dimensions of the vocal tract, and not a specific strategy of the speaker. This method has been reported to be quite successful in predicting accurate scale factors for several speaker – and therefore allowing speaker normalization – using manual extractions of formants. Here, the conditions are different; we would like to adapt the model independently on each age of our reference speaker. We expect to observe an increase in the overall length of the vocal tract, especially for the length of the pharyngeal tube.

Due to limited time, the acoustic features were obtained fully automatically, using the informations of the forced phonemic alignment. Better results would be obtained if we extracted each phoneme manually on each year, instead of computing an average on an imperfect automatic alignment. The adaptation method usually uses vowels /aeiou/, which span over the whole F1-F2 space. We had to remove the /a/ from the analysis however, due to inconsistent results for this phoneme. Table 4.1 gives some statistics about the vowels used for the adaptation. /a/ has a very large variability. This method is able to normalize the speaker with very limited data – only a few speech samples are used – but it is very sensitive to errors.

The resulting curves obtained for the two scale parameters are surprisingly good. First, despite the adverse conditions, we find consistent results over the years. We should point out that informal listening tests have indicated that the automatic alignment was far from perfect, and therefore many samples used are erroneous. Since we use only averaged values, the errors might cancel each others over the years. Second, the curves show what we were

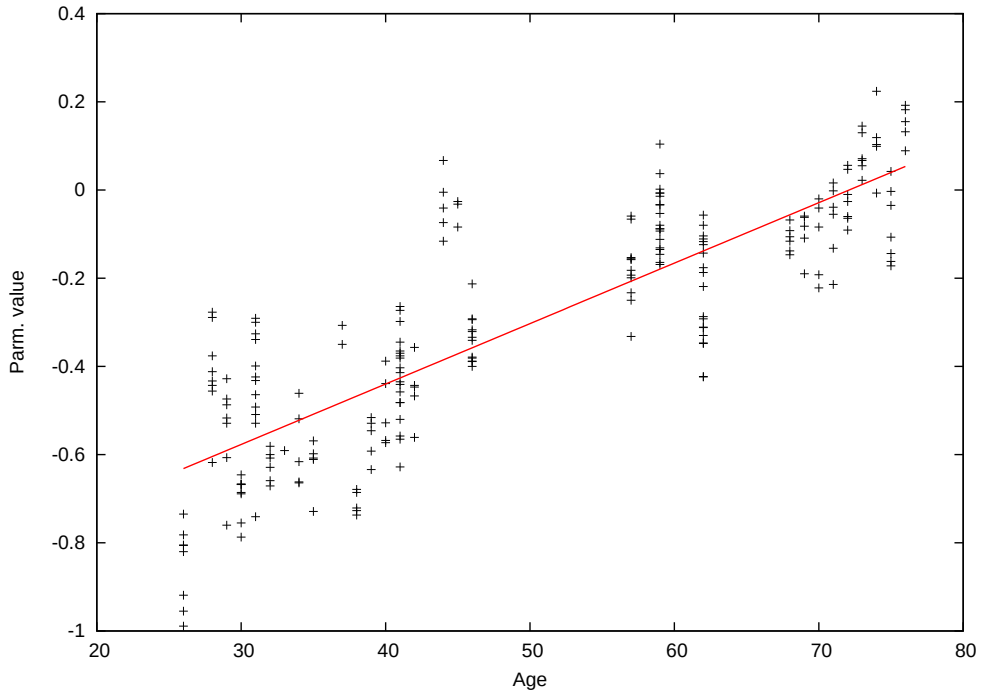

Figure 4.2: Average value for articulatory parameter P1 (Jaw) over age for the Queen.

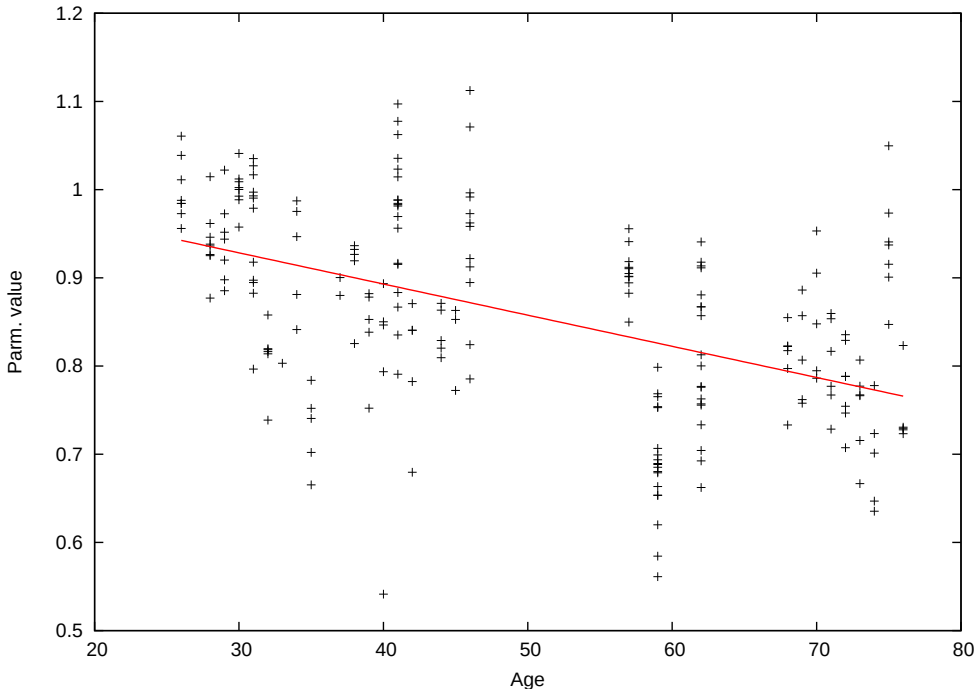

Figure 4.3: Average value and standard deviation for articulatory parameter P1 (Jaw) over age for the Queen.

Vowel	F0	F1	F2	F3
a	237.1 *53.6*	643.8 *138.6*	1285.3 *174.7*	2564.7 *303.1*
e	237.7 *52.4*	510.3 *79.5*	1929.6 *393.6*	2773.5 *279.4*
i	239.6 *54.9*	430.4 *54.0*	2004.4 *396.0*	2852.3 *241.9*
o	246.9 *50.8*	497.6 *62.5*	1497.1 *507.8*	2700.3 *269.9*
u	240.6 *49.9*	419.1 *38.8*	1548.8 *304.8*	2723.1 *219.8*

Table 4.1: Characteristics of the vowels (mean and *standard deviation* of F0...F3) used for the adaptation.

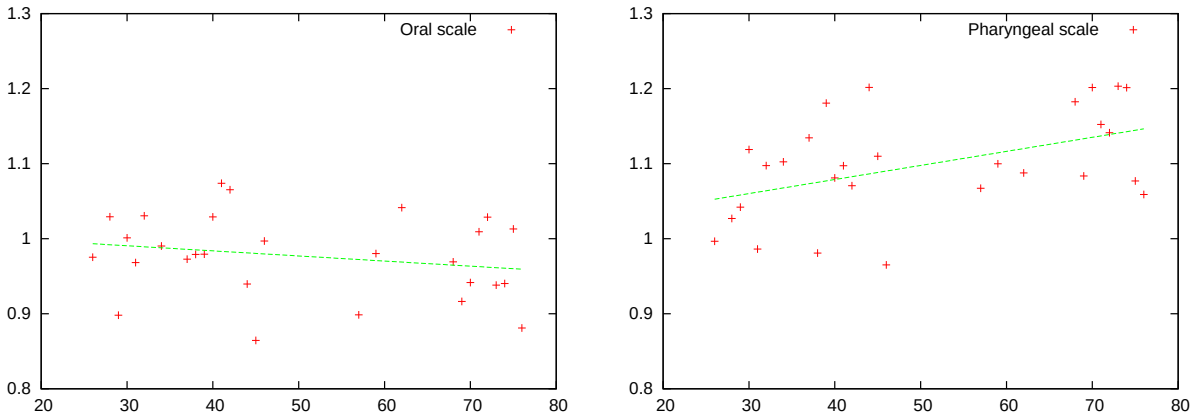

Figure 4.4: Variations of the scale parameters over age.

expecting: the length of the oral tube do not change much over the year -it even appears to be slightly decreasing-, and the length of the pharyngeal tube increases, which is consistent with the lowering of the larynx observed for aging women. The scale factors found still seem very noisy; for more consistent results, we should certainly extract manually the formants of the phonemes.

4.5 Discussion

One of the most striking observation on the trend of articulatory parameters was the increasing value of the average of P1, which indicates the position of the jaw. The jaw is well known to be the heaviest articulator: it is the most energy consuming, and also the slowest. Probably the muscles of the jaw become less flexible with age, or less dynamic, and therefore the articulation might be done through articulators that are easier to move, like the tongue. It was observed that in normal speech, the tongue movements use less that 20% of their maximum capacity; there is therefore a large place for compensation through this articulator. The tongue shape and the tongue tip parameters (P3 - P4) were with P7 the only parameters that did not see their standard deviation decrease with age.

The model adaptation performed on the Queen indicates an increase of the length of the

pharyngeal cavity with age, which is consistent with what we expected from the literature. This lengthening of the pharynx was however quite slow, and since this study was only performed on was speaker, these results need to be taken with caution.

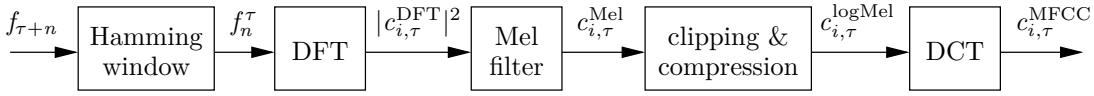

Figure 5.1: Steps to compute the Mel frequency cepstral coefficients

Chapter 5

Description of Used Features

In this chapter we describe the features we used for our age classification system (Chapter 6). First we start with spectral features, with the well known MFCCs and the formants. After that the features from the *Erlangen Prosody Module* are described. Then speaking rate, duration features and voice quality features are examined by preliminary experiments. The chapter is based on [Stei 08].

5.1 Spectral Features

The features of this group are based on the spectral short-term analysis. The Mel frequency cepstral coefficients (MFCC) are the standard features in speech recognition. They have been designed to discriminate phones and to represent what is spoken in a very compact way. Other information like the information how something is spoken should be removed. The computation of the MFCC features consists of several steps, which are illustrated in Figure 5.1. In each steps, the dimension of the feature vector is reduced: the 256 samples of the speech signal, a frame typically consists of, are finally reduced to only 12 MFCC coefficients. Formants describing the resonance frequencies of the vocal tract are another type of spectral features, which are described at the end of this section.

5.1.1 MFCC - Mel Frequency Cepstral Coefficients

The Mel Frequency Cepstral Coefficients (MFCC) have been proposed by Davis and Mermelstein [Davi 80]. In the following, the individual steps in the computation of the MFCC features are described.

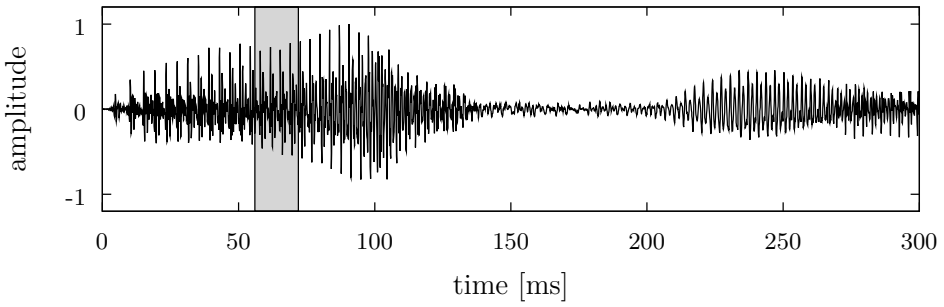

Figure 5.2: Speech signal of the word "Aibo"

DFT

Speech signals are non-stationary signals whose spectral properties change at least from one phone to another. Hence, the spectral analysis is performed on small periods of the discrete speech signal f_n, which are about 5-30 ms long. Within these so-called *windows* or *frames*, the signal can be assumed to be approximately stationary. In automatic speech recognition, a frame is typically 16 ms long corresponding to $N_\mathrm{s} = 256$ samples at a sampling rate of 16 kHz. For computational reasons, it is beneficial if N_s is a power of 2. Every 10 ms, a new frame is analyzed so that consecutive frames overlap. Figure 5.2 shows a speech signal of 300 ms duration where the word "Aibo" is spoken. A single frame of 16 ms duration is highlighted.

For each frame, the power spectrum is computed using the Discrete Fourier Transformation (DFT). The DFT assumes that the discrete and time limited signal is periodically continued. To avoid discontinuities at the beginning and the end of the frame, the amplitude of the signal is extenuated towards the borders of the window by applying a window function w_n. Let f_n^τ denote the samples of a frame that starts at sample τ after the application of the window function w_n:

$$f_n^\tau := \begin{cases} f_{\tau+n} \cdot w_n & 0 \leq n < N_\mathrm{s} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

The window function w_n is centered around $\tau + \frac{N_\mathrm{s}-1}{2}$. Various window functions such as the Hamming window, the Hann window, the Gauss window, or the Blackman window are common. For this work, the Hamming window is used which is defined as follows:

$$w_n^{\text{Hamming}} := \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N_\mathrm{s}}\right) & 0 \leq n < N_\mathrm{s} \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

Figure 5.3 shows the samples of the frame highlighted in Figure 5.2 and the window function of the Hamming window. The samples of the frame after the application of the Hamming window are shown in the left part of Figure 5.4.

Hence, the dimension of the feature vector is reduced from 256 sample to only $\frac{N_\mathrm{s}}{2}+1 = 129$ features. The right plot in Figure 5.4 shows the spectrum of the windowed speech signal that is depicted in the left part of the figure. The first local maximum represents the fundamental frequency, which is about 275 Hz in this case. The other local maxima are the harmonics, which are multiples of the fundamental frequency. Certain frequency bands are emphasized

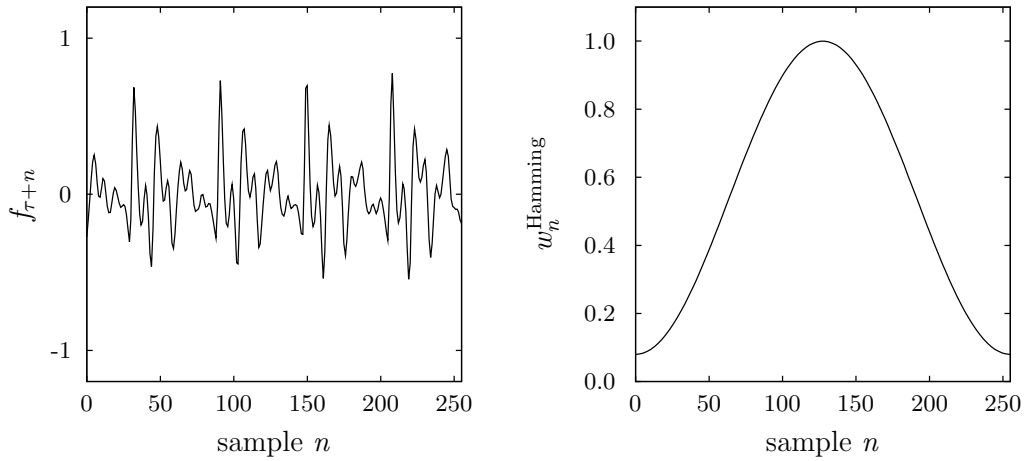

Figure 5.3: Speech frame of 256 samples starting at sample τ (left) and Hamming window (right)

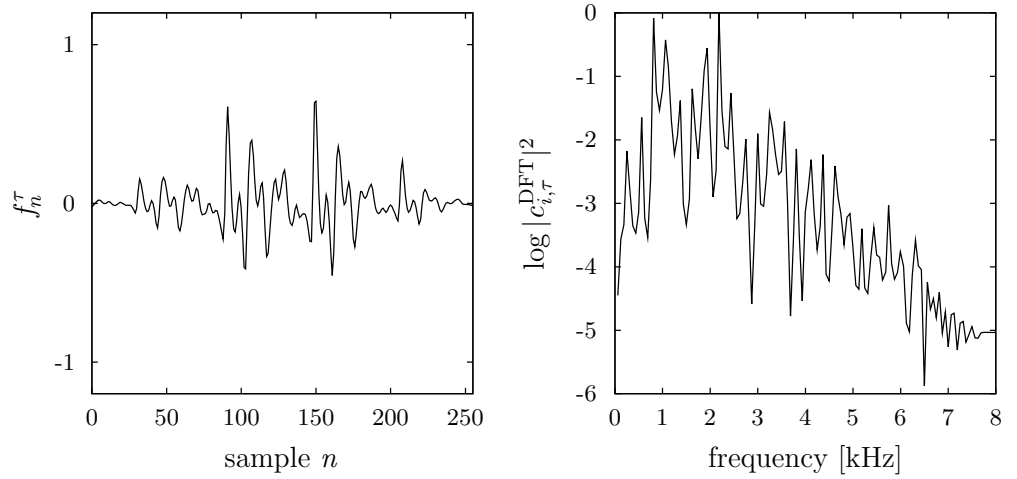

Figure 5.4: Amplitudes of the speech frame after application of the Hamming window (left) and power spectrum of the speech frame (right)

due to resonance frequencies of the vocal tract. These frequencies are known as formants and described in more detail in Chapter 5.1.2. Especially, the frequencies around the second and the third harmonic and around the sixth and the seventh harmonic are significantly higher than the surrounding harmonics due to the first and the second formant, respectively.

Mel Features

The Mel scale reflects the non-linear relationship between the frequency of a tone and the perceived pitch. In experiments by Stevens et al. , tones scattered throughout the audible range were presented at a constant loudness level of 60 dB to observers who had to adjust the frequency of a second tone until it sounded just half as high in pitch as the standard tone [Stev 37]. At a frequency of 1000 Hz, the unit of the frequency and the unit of the perceived pitch are equal: 1000 Hz $\widehat{=}$ 1000 Mel. Several quite similar equations describe this non-linear

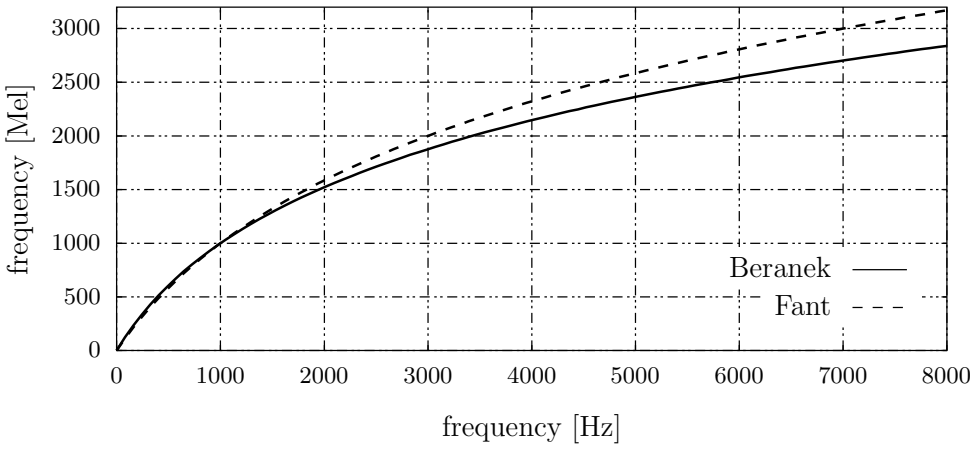

Figure 5.5: Mel scales proposed by Beranek [Bera 49] and Fant [Fant 68]

relationship between Hz and Mel found in the experiments by Stevens, e. g. the equation by Beranek [Bera 49]

$$f_{\text{Beranek}} = 1127.01048 \cdot \ln\left(1 + \frac{f_{\text{Hz}}}{700}\right) \tag{5.3}$$

or the one proposed by Fant [Fant 68]

$$f_{\text{Fant}} = \frac{1000}{\ln(2)} \cdot \ln\left(1 + \frac{f_{\text{Hz}}}{1000}\right) \ . \tag{5.4}$$

Both curves are plotted in Figure 5.5. The transformation of the frequency is approximated by applying a bank of N_{filter} filters [Riec 95]:

$$c_{i,\tau}^{\text{Mel}} = \sum_{j=1}^{N_{\text{s}}/2} w_{i,j}^{\triangle} \cdot |c_{j,\tau}^{\text{DFT}}|^2 \qquad i = 1, 2, \ldots, N_{\text{filter}} \tag{5.5}$$

The filters are designed for a higher resolution at lower frequencies. For higher frequencies, the number of filters decreases and the filters cover a wider range of frequency bands. This models the decreasing frequency resolution for higher frequencies of the human auditory system. The weighted summation of adjacent frequency coefficients removes the harmonic structure of the spectrum as it can be seen from the Mel spectrum shown in the left figure of Figure 5.7. Furthermore, the number of coefficients is reduced from $N_{\text{s}}/2 + 1 = 129$ to N_{filter}. For this work, $N_{\text{filter}} = 22$ triangular filters are used. They are depicted in Figure 5.6.

Approximately, the Mel spectrum coefficients $c_{i,\tau}^{\text{Mel}}$ are distributed log-normally. For a classification with Gaussian mixture models, this is not favorable if the number of mixtures is low. Hence, the Mel spectrum coefficients are compressed by taking the logarithm. In order to prevent numerical problems if the logarithm is taken of values that are close to zero, the Mel spectrum coefficients are clipped to the interval $[\epsilon; 1]$ prior to the compression:

$$c_{i,\tau}^{\text{normMel}} = \begin{cases} \frac{c_{i,\tau}^{\text{Mel}}}{\max\limits_{j} c_{j,\tau}^{\text{Mel}}} & c_{i,\tau}^{\text{Mel}} > \epsilon \max\limits_{j} c_{j,\tau}^{\text{Mel}} \\ \epsilon & \text{otherwise} \end{cases} \qquad i = 1, 2, \ldots, N_{\text{filter}} \tag{5.6}$$

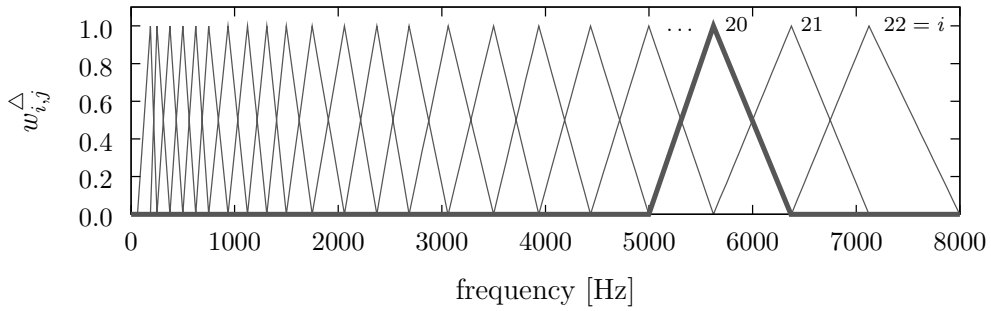

Figure 5.6: Mel filter bank consisting of 22 triangular filters covering a frequency range of 80 Hz to 8 kHz. The coefficients $w^{\triangle}_{20,j}$ of the 20th filter are highlighted

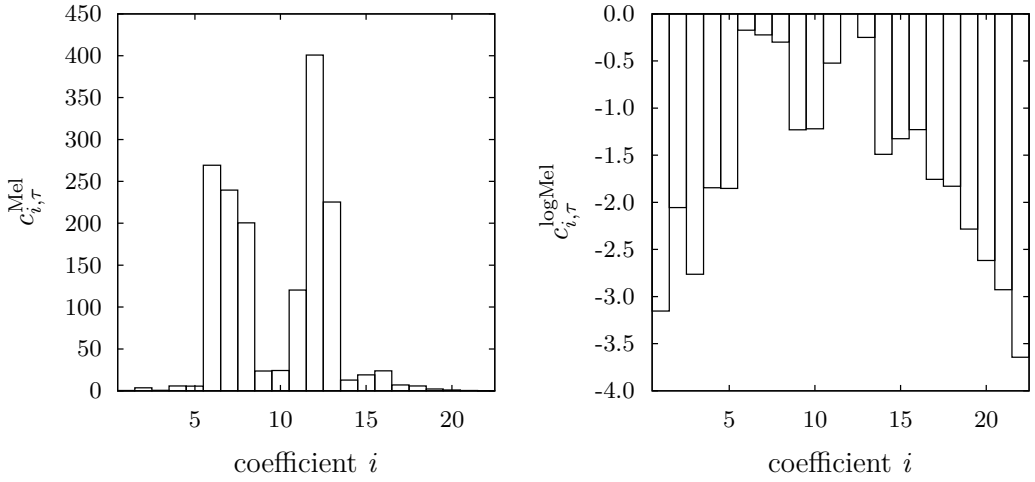

Figure 5.7: Mel spectrum (left) and log-Mel spectrum of the speech frame (right)

In the experiments, ϵ is set to 0.000001. Finally, the compressed Mel spectrum coefficients $c^{\text{logMel}}_{i,\tau}$ are obtained by:

$$c^{\text{logMel}}_{i,\tau} = \log_{10}\left(c^{\text{normMel}}_{i,\tau}\right) \qquad i = 1, 2, \ldots, N_{\text{filter}} \tag{5.7}$$

The compressed Mel spectrum is shown in the right part of Figure 5.7.

MFCC Features

In the last step to extract the MFCC features the Discrete Cosine Transformation (DCT) is applied to the log-Mel spectrum of the signal:

$$c^{\text{MFCC}}_{i,\tau} := \sqrt{\frac{2}{N_{\text{filter}}}} \sum_{j=1}^{N_{\text{filter}}} c^{\text{logMel}}_{j,\tau} \cdot \cos\left(\frac{i(j-\frac{1}{2})\pi}{N_{\text{filter}}}\right) \quad i = 0, 1, \ldots, N_{\text{filter}} - 1 \tag{5.8}$$

The DCT is a Fourier related transformation that can be applied to real data with even symmetry. The output of the DFT of the speech signal meets these requirements. Hence, the spectrum of the log-Mel spectrum is computed. The domain is called *cepstrum*, a term

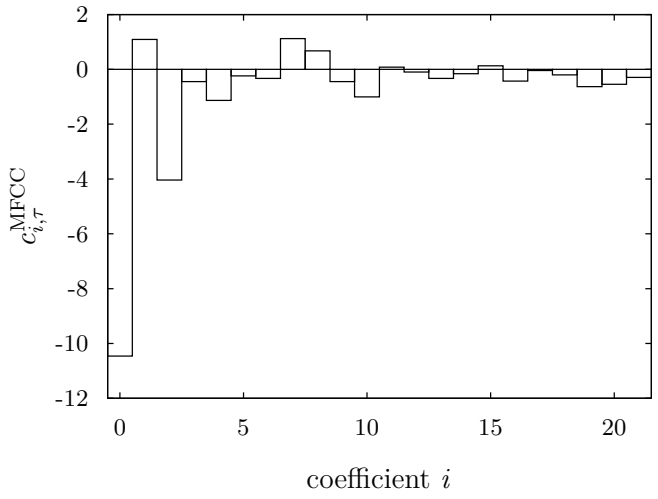

Figure 5.8: Mel frequency cepstral coefficients (MFCC)

made up by reversing the letters of the first syllable of 'spectrum'. Only the first 12 MFCC coefficients are taken discarding the coefficients that represent higher frequencies. The MFCC coefficients are depicted in Figure 5.8.

The Discrete Cosine Transformation decorrelates the log-Mel coefficients similar to the principal component analysis (PCA) [Ahme 74] but with the advantage of a constant transformation matrix and without the need to compute the eigenvectors of the data.

The first MFCC coefficient $c_{0,\tau}^{\mathrm{MFCC}}$ is substituted by the logarithm of the short-time energy c_{τ}^{en} defined as the sum of all Mel spectrum coefficients:

$$c_{\tau}^{\mathrm{en}} = \log_{10}\left(\sum_{i=1}^{N_{\mathrm{filter}}} c_{i,\tau}^{\mathrm{Mel}}\right) \tag{5.9}$$

In order to reduce the impact that changes of the environmental conditions such as noise, room, microphone, or speaker characteristics have on the MFCC features, techniques like the *cepstral mean subtraction* (CMS) are applied where a pre-computed mean of the MFCC feature vector is subtracted. An extension, which is used in our implementation, is the *dynamic adaptive cepstral subtraction* (DACS) where the pre-computed mean is updated while new frames are processed. Only frames that actually contain speech are used for the adaptation of the mean.

The MFCC coefficients $c_{i,\tau}^{\mathrm{MFCC}}$ are called *static* features as they describe the spectral properties within one frame where the signal is approximately stationary. The feature vector is extended by *dynamic* features, which describe the behavior of the static features over time. For this purpose, the first and sometimes also the second derivative of the static features are calculated. These features are often called Δ and $\Delta\Delta$ features, respectively. For this work, only Δ features are investigated. The first derivative is approximated by the slope of the regression line [Furu 86] that is fitted to the MFCC feature vectors of five consecutive frames:

$$c_{i,\tau}^{\Delta\mathrm{MFCC}} = \frac{\sum_{j=-2}^{2} j \cdot c_{i,\tau+j}^{\mathrm{MFCC}}}{\sum_{j=-2}^{2} j^2} \tag{5.10}$$

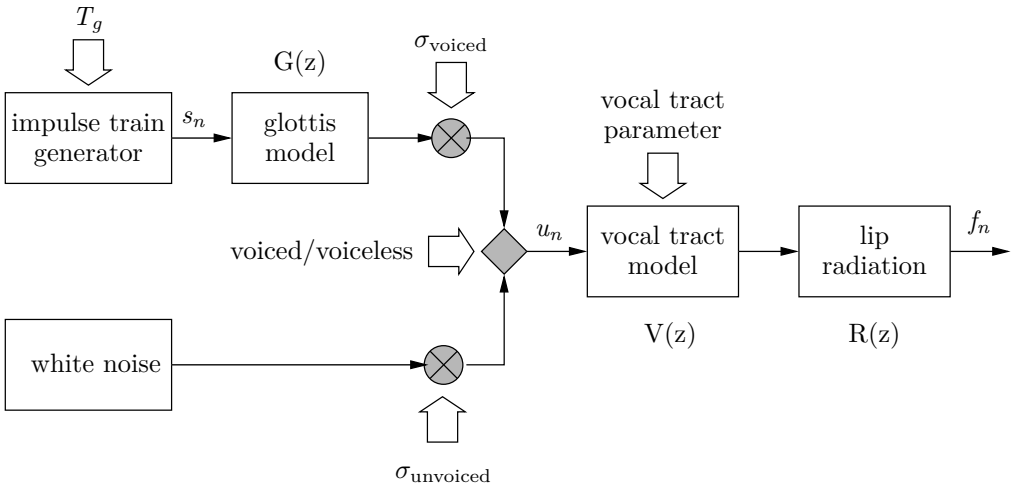

Figure 5.9: Fant's Source-filter model after [Schu 95]

5.1.2 Formant Based Features

Fant's source-filter model [Fant 60] illustrated in Figure 5.9 models the process of speech production as a series of linear, time invariant systems. The discrete speech signal can be obtained by the convolution $f_n = u_n \star v_n \star r_n$. The following equation holds for the z-transform of the speech signal:

$$F(z) = U(z) \cdot V(z) \cdot R(z) \tag{5.11}$$

The complete transmission function for voiced sounds is $H(z) = \sigma_{\text{voiced}} \cdot G(z) \cdot V(z) \cdot R(z)$. $G(z)$ is the z-transform of the glottis model, $V(z)$ the one of the vocal tract and $R(z)$ models the radiation at the lips.

In order to model the resonance characteristics of the vocal tract, the vocal tract is modeled in a simplified way by an acoustic tube of the length L consisting of M cylindrical segments as illustrated in Figure 5.10. The nasal tract and losses at the wall of the vocal tract are not modeled. All segments have the same length $l = L/M$ but different cross sectional areas A_i, $1 \leq i \leq M$. The typical length of the vocal tract is about $L = 170\,\text{mm}$ for adults. In the direction of the tube, a planar propagation of the signal can be assumed since the length of the cylindrical segments is far below the wave length of speech signals [Schu 95]. The acoustic flow in the forward and the backward direction can be computed iteratively from the reflection coefficients

$$k_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}, \quad 0 \leq i \leq M \tag{5.12}$$

The area A_0 of the "outside world" cylinder in front of the lips is set to infinity; then k_0 is 1. The area of the terminator at the glottis does not affect the resonance characteristics and can be chosen arbitrarily. The propagation of the signal is disturbed only at equidistant points of time due to the change of the diameter of the tube at the transition from one cylinder to another. Hence, a simple term results for the z-transform of the vocal tract:

$$V(z) = \frac{\prod_{i=0}(1 + k_i)}{1 - \sum_{i=1}^{M} a_i z^{-i}} \tag{5.13}$$

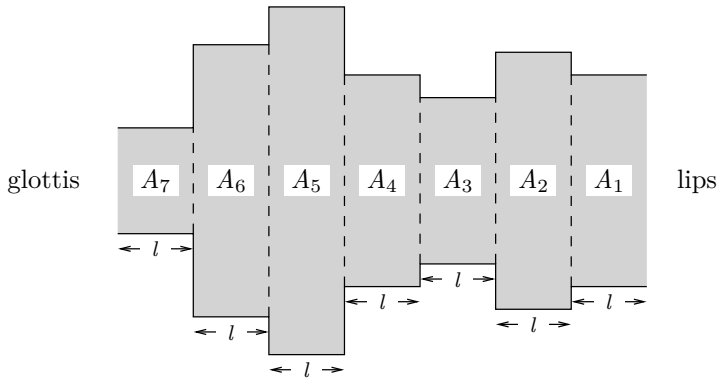

Figure 5.10: Vocal tract model: acoustic tube without loss consisting of cylindrical segments of equal length after [Schu 95]

The polynomial coefficients $a_i = a_i^{(M)}$ in the denominator can be computed iteratively:

$$a_i^{(m)} = \begin{cases} 1 & i = 0 \\ a_i^{(m-1)} + k_m \, a_{m-i}^{(m-1)} & i < 1 < m \\ k_m & i = m \end{cases} \tag{5.14}$$

The function $V(z)$ has $M/2$ pairs of complex conjugate poles where the polynomial in the denominator is equal to zero:

$$1 - \sum_{i=1}^{M} a_i z^{-i} = \prod_{i=1}^{M/2} \left(1 - 2e^{-c_i T} \cos(b_i T) z^{-1} + e^{-2c_i T} z^{-2}\right) \tag{5.15}$$

These poles are the resonances of the vocal tract well-known as *formants*. They are characterized by their center frequencies $F_i = b_i/(2\pi)$ and their bandwidths $B_i = c_i/(2\pi)$. The formants characterize the current shape of the vocal tract while a phone is being produced. They are independent of the perceived pitch. Yet, they do depend on the length of the vocal tract and hence, depend on the age and the gender of the speaker. Significant differences in the position of the first two formants between adults and children have been found [Stem 05]. Nevertheless, the first two formants are sufficient to identify vowels. Algorithms that determine the formants by finding the poles of $V(z)$ are called *root extraction methods*.

Other algorithms, called *spectral peak picking methods*, extract the local maxima of a smoothed spectrum such as the one obtained by linear prediction coding (LPC). The LP spectrum is shown in Figure 5.11. LPC assumes that the samples of a stationary period of the signal can be predicted by a linear combination of the preceding N_{LP} samples:

$$\hat{f}_n^\tau = -\sum_{j=1}^{N_{\mathrm{LP}}} \alpha_j \cdot f_{n-j}^\tau \tag{5.16}$$

N_{LP} is called the prediction order; α_j, $1 \le j \le N_{\mathrm{LP}}$, denote the LP coefficients. There will be a deviation between the predicted value $\hat{f}_n^\tau$ and the actual value f_n^τ:

$$e_n^\tau = f_n^\tau - \hat{f}_n^\tau = \sum_{j=0}^{N_{\mathrm{LP}}} \alpha_j \cdot f_{n-j}^\tau, \quad \text{with} \quad \alpha_0 = 1 \tag{5.17}$$

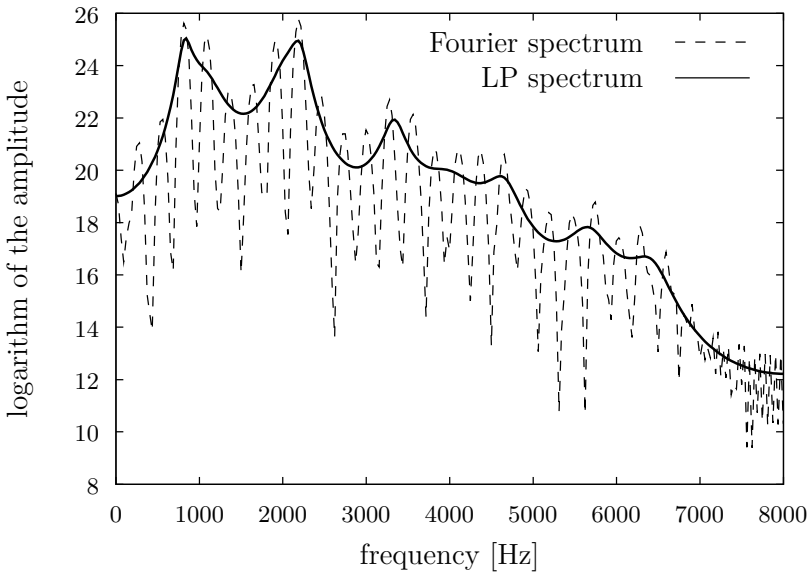

Figure 5.11: Log-Fourier and log-LP spectrum ($N_{\mathrm{LP}} = 20$) of the speech frame

The LP coefficients $\alpha_1, \ldots, \alpha_{N_{\mathrm{LP}}}$ are determined such that the accumulated squared error

$$\epsilon_{\mathrm{LP}} = \sum_{j=N_{\mathrm{LP}}}^{N_{\mathrm{s}}-1} (e_n^\tau)^2 = \sum_{j=N_{\mathrm{LP}}}^{N_{\mathrm{s}}-1} \left(\sum_{j=0}^{N_{\mathrm{LP}}} \alpha_j \cdot f_{n-j}^\tau \right)^2 \tag{5.18}$$

is minimized. Again, N_{s} is the number of samples in the speech frame. The error ϵ_{LP} can be rewritten in a more compact form as a linear combination of quadratic functions:

$$\epsilon_{\mathrm{LP}} = \sum_{j=0}^{N_{\mathrm{LP}}} \sum_{k=0}^{N_{\mathrm{LP}}} \alpha_j \phi_{jk}^\tau \alpha_k \quad \text{with} \quad \phi_{jk}^\tau = \sum_{n=0}^{N_{\mathrm{s}}-1} f_{n-j}^\tau \, f_{n-k}^\tau \tag{5.19}$$

Hence, ϵ_{LP} has a unique minimum which can be found by setting the partial derivatives $\partial \epsilon_{\mathrm{LP}} / \partial \alpha_k = \sum_{j=0}^{N_{\mathrm{LP}}} \alpha_j \phi_{jk}^\tau$ to zero. This results in a system of N_{LP} linear equations:

$$\sum_{j=1}^{N_{\mathrm{LP}}} \alpha_j \phi_{jk}^\tau = -\phi_{jk}^\tau, \quad 1 \leq k \leq N_{\mathrm{LP}} \tag{5.20}$$

Instead of solving this system, the LP coefficients can be determined in a faster way using the *covariance method* or the *autocorrelation method*. The first one in based on a Cholesky decomposition of the symmetric matrix $\boldsymbol{\phi}^\tau$, the latter one uses the autocorrelation function r to compute the components of $\boldsymbol{\phi}^\tau$:

$$\phi_{jk}^\tau = r_{|j-k|}^\tau \tag{5.21}$$

and benefits from the form of $\boldsymbol{\phi}^\tau$, which is a Toeplitz matrix (constant elements on each descending diagonal from left to right). In this case, the system of equations can be solved with the Levinson-Durbin recursion [Levi 47, Durb 60]. In order to compute the LP spectrum,

the LP coefficients are zero padded before applying the Discrete Fourier Transformation. Finally, the local maxima of the spectrum are determined. The number of local maxima strongly depends on the prediction order N_{LP}. If the order is too low, one local maximum models more than one formant. If N_{LP} is too high, the local maxima model the harmonic structure. Generally, good results are obtained if N_{LP} is set to $f_{\mathrm{s}} + 4$; f_{s} is the sampling frequency given in kHz.

In the experiments, the first three formants *without* their bandwidths are used. They are extracted using the formant extractor of ESPS, which is incorporated in the software WaveSurfer. The algorithm is based on the root extraction method.

5.2 Features of the *Erlangen Prosody Module*

The Erlangen Prosody Module has been originally designed to detect prosodic events such as phrase boundaries, phrase accents, and sentence mood in order to improve the automatic processing of speech [Warn 03, Gall 02, Noth 02, Komp 97, Kies 97, Noth 91, Noth 88]. The Erlangen Prosody Module has been an integral component of the German Verbmobil project [Batl 00b, Batl 00a, Noth 00] and the SmartKom project [Zeis 06].

The features of the Erlangen Prosody Module model the contour of the fundamental frequency and the short-term energy, aspects of temporal lengthening of words, and the duration of pauses. In total, the Erlangen Prosody Module computes 100 prosodic features for each word: 26 F_0 based features, 33 energy based features, 33 duration based features, and 8 features based on pauses.

Besides its main purpose to calculate prosodic features, the Erlangen Prosody Module can also be used to calculate some non-prosodic features such as jitter and shimmer features, which are described in the section on voice quality features (s. Chapter 5.2.4). In the following subsections, a detailed description of the prosodic features is given.

5.2.1 F_0 Based Features

The F_0 based features model the contour of the (logarithmic) fundamental frequency as it is illustrated in Figure 5.12. In detail, the contour is described by the slope of the regression line, the error that occurs if the contour is approximated by this line, the maximum and the minimum of the fundamental frequency, and the F_0 onset and the F_0 offset, i. e. the F_0 values at the first voiced frame and the last voiced frame, respectively. Furthermore, the average of the F_0 values within one word is included. The position of both extrema and the positions of the on- and the offset are temporal measures specifying the distance from a given reference point that is defined as the end of the current word. In the experiments, these temporal features are treated separately or in combination with the duration based features described in Chapter 5.2.3.

The feature vector for the word under consideration is extended by the features of the words constituting the left and the right context. A context of at most two words to the left and two words to the right is considered since former experiments have shown that larger context sizes do not improve the classification performance [Batl 00a]. Reasons that are pointed out are either the fact that a larger context does not contain relevant information

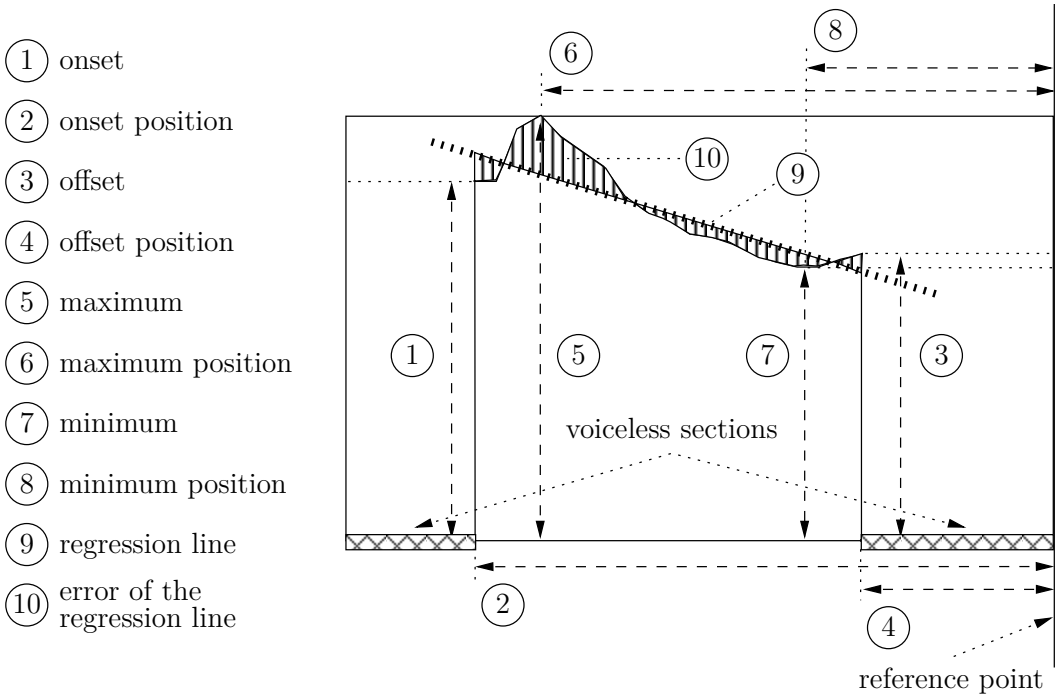

Figure 5.12: Features of the Erlangen Prosody Module after [Buck 99]

feature	context				
	−2	−1	0	+1	+2
maximum		•	•	•	
minimum		•	•	•	
mean		•	•	•	
	whole turn				
onset			•	•	
offset		•	•		
regression coefficient		•	•	•	
		•			
	•			•	
regression error		•	•	•	
		•			
	•			•	

Figure 5.13: 26 local F_0 based features and their context of computation

to model the local events, or the rather limited size of the training data that has been used in the Verbmobil project. Table 5.13 shows which features are calculated for which context [Hube 02, Kies 97]. The slope of the regression line and the corresponding approximation error is also calculated for speech segments covering two words. Furthermore, the mean F_0 value for the whole turn is included. In total, 26 F_0 based features are extracted for each word.

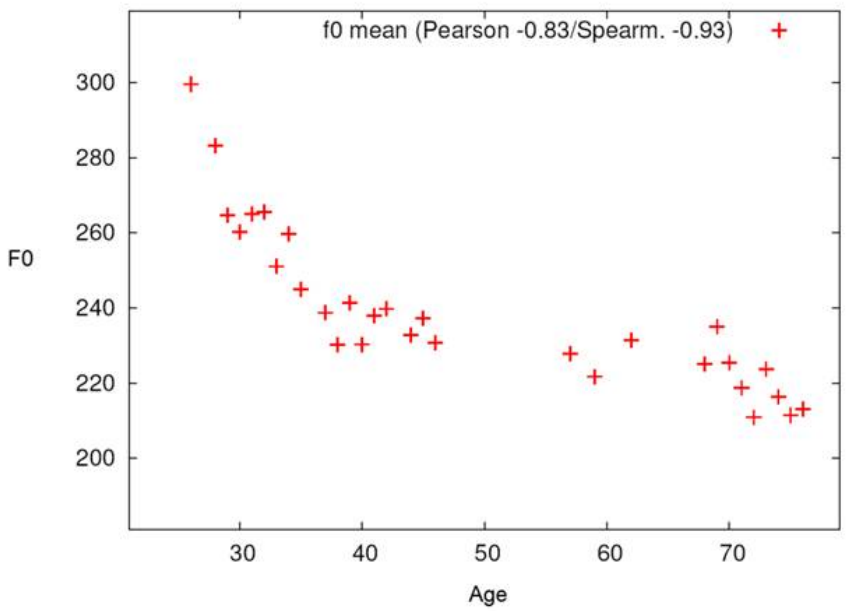

Figure 5.14: Averaged F_0 of the recordings of Queen Elizabeth II.

For a first look on the F_0 we calculated the per-year-averaged F_0 on the Queen recordings. The results are plotted in Figure 5.14. The age-correlation is -0.83/-0.93 (Pearson/Spearman).

5.2.2 Energy Based Features

Similar to the F_0 based features, the energy based features model the contour of the short-term energy of each frame (frames of 16 ms duration, time shift of 10 ms). Certain statistics like the minimum, which is always zero or close to zero, and the on- and the offset do not make any sense and are excluded [Kies 97]. In contrast, the position of the minimum may make sense very well. Again, the positions of the extrema are treated as duration features. In addition, the energy of the whole word is included: once as its absolute value and once in a normalized form.

The normalized energy of a word is based on the work of Wightman [Wigh 92]. The energy factor τ_{en} specifies how much louder or softer the speaker produces the words in an interval I compared to an average speaker.

$$\tau_{\text{en}}(I) := \frac{1}{\#I} \sum_{w \in I} \frac{\text{en}(w)}{\mu_{\text{en}}(w)} \tag{5.22}$$

$\text{en}(w)$ denotes the energy of the word w. The statistics $\mu_{\text{en}}(w)$ and $\sigma_{\text{en}}(w)$ are the average energy of the word w produced by an average speaker and the corresponding standard deviation, respectively. If the frequency of the given word is too small to obtain robust estimates of the statistics, they can be approximated based on the energy statistics of the syllables or phonemes that the word consists of. The energy factor τ_{en} is added to the feature

feature	context				
	−2	−1	0	+1	+2
maximum		•	•	•	
mean		•	•	•	
	•			•	
	•		•		
regression coefficient		•	•	•	
		•			
	•			•	
regression error		•	•	•	
		•			
	•			•	
absolute energy of a word		•	•	•	
	•			•	
normalized energy ζ_{en}		•	•	•	
	•			•	
τ_{en}	whole turn				

Figure 5.15: 33 local energy based features and their context of computation

vector and is constant for all words within one turn. Furthermore, the factor τ_{en} is used to scale the expected energy $\mu_{\text{en}}(w)$ of the word w in order to adapt the expected energy to the energy level of the whole turn. The difference $\text{en}(w) - \tau_{\text{en}}(I)\mu_{\text{en}}(w)$ in the numerator of Equation 5.23 is the deviation of the energy of the current word from its expected energy. In order to get rid of the speech sound dependent variation, this deviation is normalized with the standard deviation $\sigma_{\text{en}}(w)$ which is also scaled by the factor τ_{en}. The resulting feature $\zeta_{\text{en}}(J, I)$ for single words – in this case the interval J consists of only the current word – or for larger contexts is defined as follows:

$$\zeta_{\text{en}}(J, I) := \frac{1}{\#J} \sum_{w \in J} \frac{\text{en}(w) - \tau_{\text{en}}(I)\mu_{\text{en}}(w)}{\tau_{\text{en}}(I)\sigma_{\text{en}}(w)} \tag{5.23}$$

As for F_0 based features, features describing the context of the word are included resulting in 33 energy based features. Table 5.15 illustrates which features are calculated for which context.

5.2.3 Duration Based Features

Duration based features model aspects of temporal lengthening of words or segments. Besides the absolute duration of a word, two normalized forms are added to the feature vector. The first normalization is rather simple and normalizes the duration of a word by the number of syllables the word consists of. The second normalization is along the same lines as for the energy normalization. The factor τ_{dur} is the speaking rate. For its computation, only 'en' has to be substituted by 'dur' in Equation 5.22. $\text{dur}(w)$ denotes the duration of the word w. The statistics $\mu_{\text{dur}}(w)$ and $\sigma_{\text{dur}}(w)$ are the average duration of word w produced by an average speaker and the corresponding standard deviation, respectively.

feature	context				
	−2	−1	0	+1	+2
absolute duration of a word		•	•	•	
	•			•	
duration of a word normalized with the number of syllables		•	•	•	
	•			•	
normalized duration ζ_{dur}		•	•	•	
		•			
	•			•	
speaking rate τ_{dur}	whole turn				
position of the F_0 maximum		•	•	•	
position of the F_0 minimum		•	•	•	
position of the F_0 onset			•	•	
position of the F_0 offset		•	•		
position of the energy maximum		•	•	•	
position of the energy minimum		•	•	•	

Figure 5.16: 33 local duration based features and their context of computation

Table 5.16 shows which features are computed for which contexts resulting in a 17-dimensional feature vector. In addition, there are 16 features describing the positions of the F_0 and energy extrema as it has been mentioned in the previous two sections. All in all, there are 33 duration based features.

The F_0 and energy based features extracted by the Erlangen Prosody Module are combined in a feature vector of dimension 187, with one feature vector per voiced segment.

5.2.4 Voice Quality Features

Voice quality features characterize the source signal which emerges from the oscillation of the vocal cords. The source signal can be estimated by inverse filtering of the speech signal canceling the effects of the vocal tract.

Jitter and Shimmer

The term *jitter* denotes cycle-to-cycle variations of the fundamental frequency. Here, an approximation of the first derivative of the fundamental frequency is used:

$$\text{jitter}(i) = \frac{|F_0(i+1) - F_0(i)|}{F_0(i)} \tag{5.24}$$

These variations are not perceived as changes of the pitch but as changes of the voice quality. Along the same lines, the term *shimmer* denotes variations of the energy from one cycle to another:

$$\text{shimmer}(i) = \frac{|\text{en}(i+1) - \text{en}(i)|}{\text{en}(i)} \tag{5.25}$$

Cycle-to-cycle variations require that the fundamental frequency is calculated for consecutive periods and not as the average F_0 for a whole frame of constant length. The PDDP (**p**eriod **d**etection by means of **d**ynamic **p**rogramming) algorithm described in [Kies 97] is applied. In a first step, segments of voiced speech are located and estimates of the average fundamental frequency in these segments are computed. Then, the possible candidates for the period boundaries are located at positive zero crossings of the signal. Characteristics like the integral or the extrema in the segments from one positive zero crossing to the next one are computed and can be used to reduce the number of candidates by eliminating irrelevant candidates like those with negative integral. Within a voiced segment, hypotheses of periods are generated: Each positive zero crossing can be the starting point of several periods. The period length is limited to an interval that is defined by the average fundamental frequency estimated for the whole voiced segment and the maximal relative deviation from this value that is permitted. Dynamic programming (DP) is used to find the optimal path in the graph of hypotheses. Two types of cost functions are defined: cost functions that characterize the period itself and cost functions that characterize the similarity of consecutive periods. Several heuristic cost functions are defined in [Kies 97]. They are combined to a single cost function by an artificial neutral network classifier which has been trained on manually labeled positions of periods. For a more detailed description of the algorithm, please see [Kies 97].

Features for words are obtained by averaging the jitter and shimmer values over all detected periods within the given word. Additionally, the feature vector is extended by the standard deviation of the jitter and shimmer values.

Age Correlation

Our preliminary experiments on the longitudinal data of Alistair Cooke's speech show for the per year averaged jitter an 0.66 age correlation (Figure 5.17). In contrast the the shimmer feature does not show a trend (0.06 correlation).

5.3 Speaking Rate and Plosive Vowel Transition Duration

In this section we present our analysis of speaking rate and plosive vowel transition duration we tried to explore. We use forced time alignment provided by Speech Recognition team and did analysis on acoustic features such as speaking rate, silence percentage and duration features. In a previous chapter 3, we saw that there is no systematic error over age in automatically extracted formant values. So we can use automatically extracted formant values reliably for further analysis. Using formant values along with additional information, we tried to find good features for age classification.

5.3.1 Speaking Rate

As a person gets older, it is more effort for him to articulate different phones which results in low speaking rate. We define speaking rate as number of phones uttered per second. We do not consider silences and short pauses while counting the number of phones. We carried

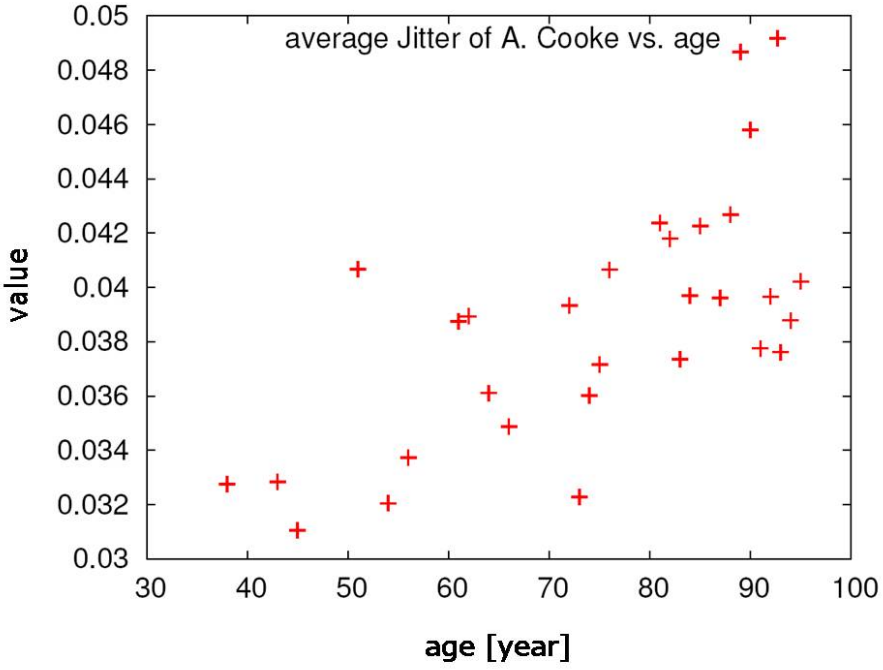

Figure 5.17: Averaged Jitter of the recordings of Alistair Cooke.

speaking rate experiments for UF-VAD males and females data and for the queen data. For lateral data, we confirmed that speaking rate is a good feature for age classification with a correlation of -0.74. We can see from Figure 1, a clear trend $young > middle > old$.

However, interestingly, for the queen data, as shown in Figure 2, we observed that the speaking rate increases till she was 50 years old and then it decreases. We speculate that this might be the effect of professionalism. Being young, she might be somewhat uncertain or nervous, which makes her speak slowly. We also observed that there is high variation in the speaking rate from 26-50. However, after 50, there is less variation showing that she is stable with her speaking rate.

5.3.2 Plosive Vowel Transition Duration

The formal definition of plosive-vowel transition duration is the time interval between the release of the oral constriction for production of a plosive and the initiation of glottal pulsing for the vowel that follows is called the plosive-vowel transition. As a person gets older the duration should increase as it is some effort for him to make such transitions. For example, consider plosive **'g'** followed by vowel **'a'**. Figure 3 explains how the formant frequencies change from **'g'** to **'a'**. The time taken for this transition is called the transition duration.

To find out the exact transition point is a hard problem. The transition point is determined for each of F1, F2 and F3 and then the decision is made about which transition point should be considered for the combination of plosive-vowel. We observed that individual plosive-vowel combinations do not reveal much information about age. So we made classes of these combinations. for certain phonetic contexts we saw a slight duration increase with age. Figure 4 demonstrates plosive-vowel duration for UF-VAD -males data and eh class. We observed that the duration slightly increases over age. The general trend is $Young < Middle < Old$ and the variability also very slightly increases over age $Young < Middle < Old$.

But for the Queen, as shown in Figure 5, we can see that the duration slightly increases over age. The general trend is $Young < Middle < Old$ and the variability decreases over age $Young > Middle > Old$. This again be might be because of the professionalism.

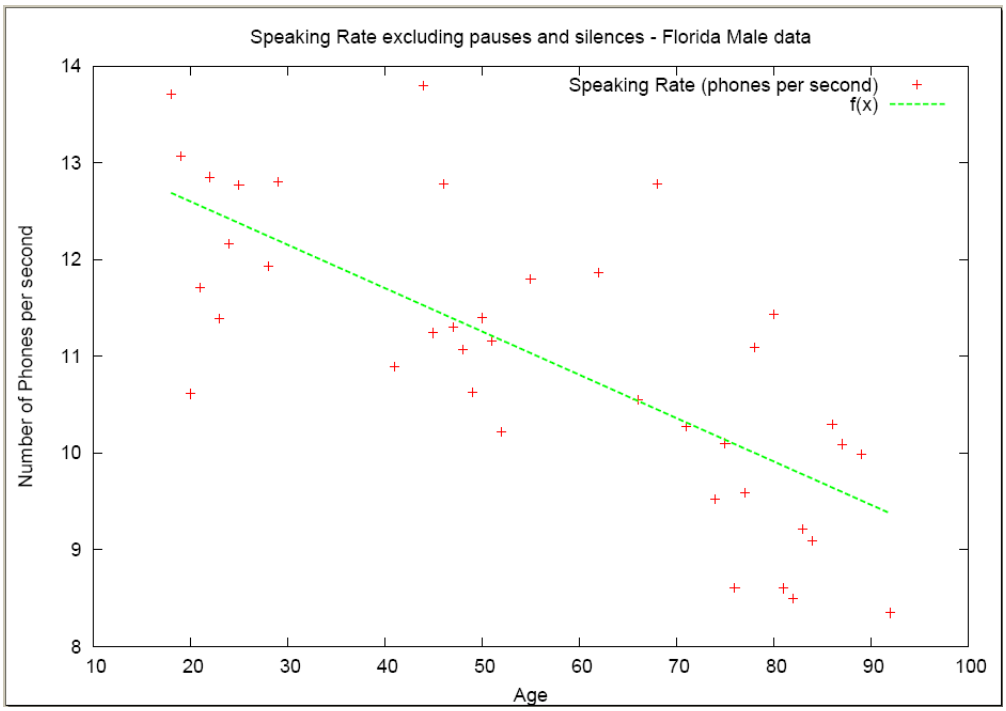

Figure 5.18: Speaking Rate (UF-VAD - male's data)

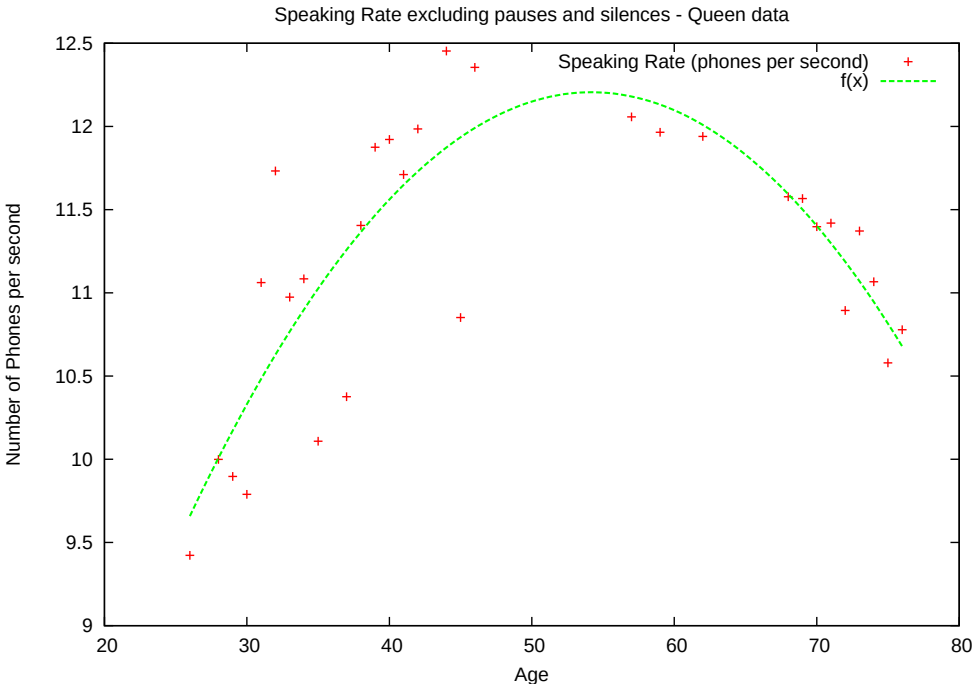

Figure 5.19: Speaking Rate (The Queen)

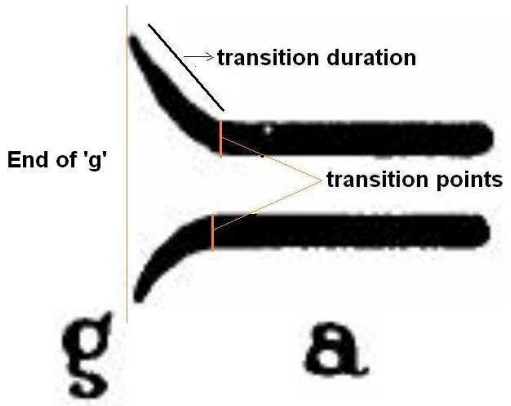

Figure 5.20: Plosive-Vowel Transition duration

5.3.3 Pause Percentage

Pause percentage is the percentage of pauses in the whole utterance. For example, if the whole utterance is 10 second long and the the total time taken by pauses is 2 seconds then the pause percentage will be 20%. Pause percentage feature seems to be slightly increasing with age with a weak correlation of 0.55. Figure 6 shows the trend *Young* < *Middle* < *Old* for UF-VAD males data.

5.3.4 Conclusion

Here is the table that summarizes the results.[1]

Variation of features with increasing age

	UF-VAD		The Queen
	Male	Female	
Speaking Rate	decrease	decrease	U shaped curve
Plosive-vowel Duration	slight increase	no trend	very slight increase
Pause Percentage	increase	slight increase	U shaped curve

[1]We know while doing the analysis we used forced time alignment and formants which might be erroneous resulting in wrong conclusions.

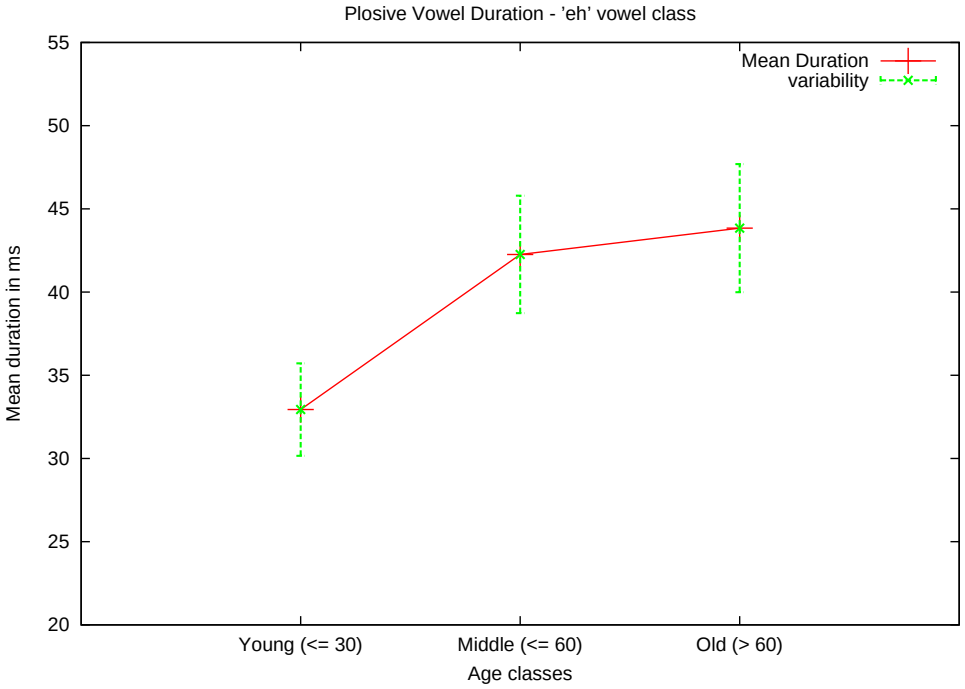

Figure 5.21: Plosive-vowel transition duration - UF-VAD males data 'eh' class

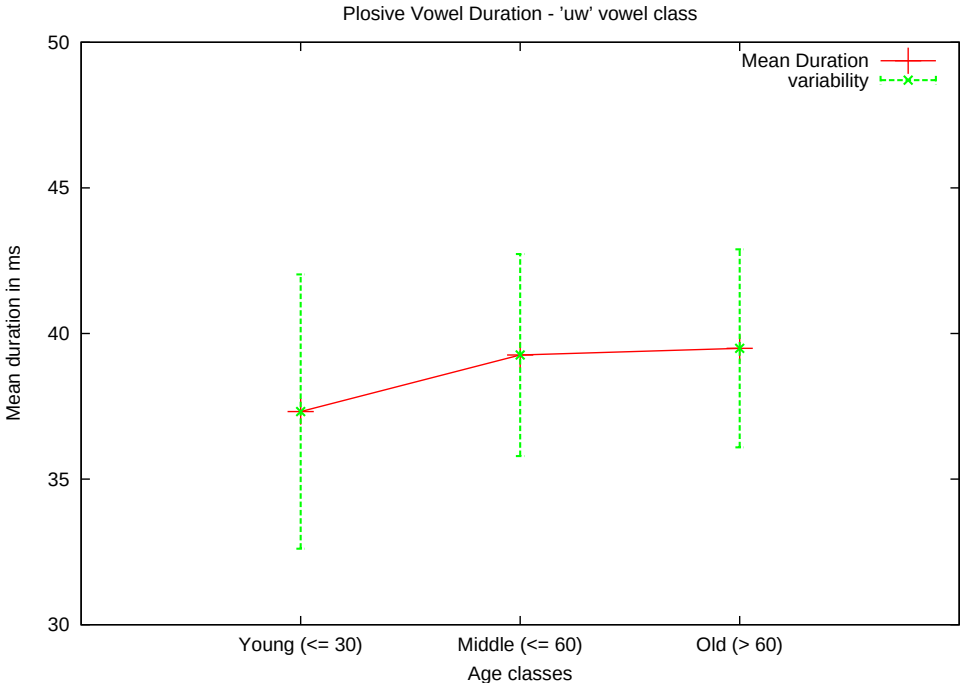

Figure 5.22: Plosive-vowel transition duration - The Queen 'uw' class

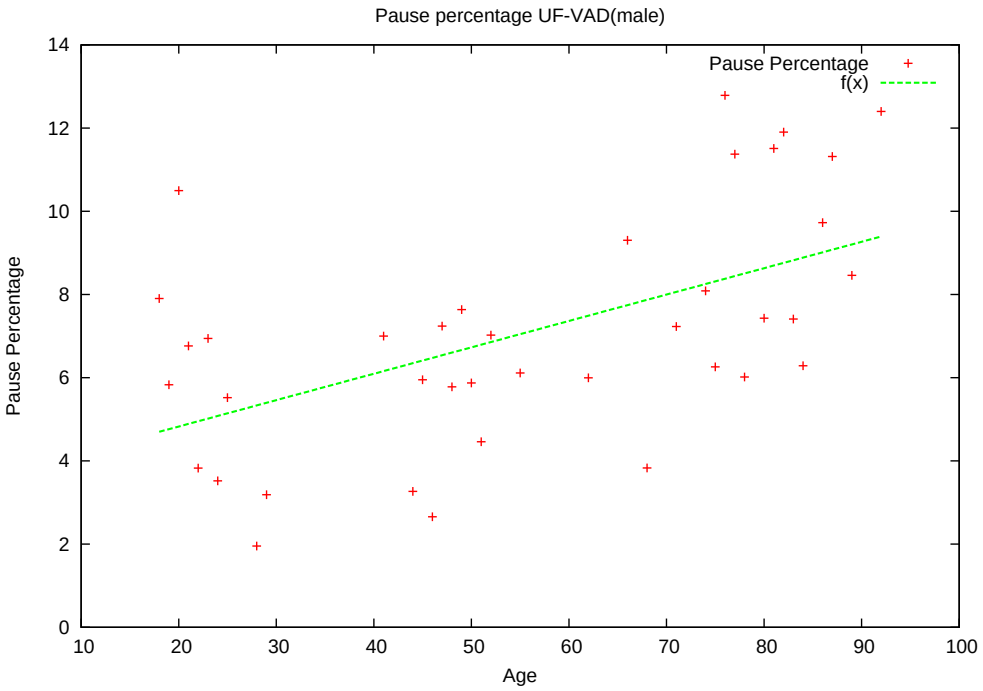

Figure 5.23: Silence Percentage - UF-VAD males data

Chapter 6

Age Prediction

We created a system that would estimate the age of a speaker from speech. Using our features, we created a training model using Gaussian Mixture Models (GMM) and evaluated using our testing set with Support Vector Regression (SVR). Since we did not have a large amount of speech data, for training we first created a Universal Background Model (UBM) on the entire dataset, and then adapted individual age/speaker models using Maximum A Posteriori (MAP) estimation. This process is explained in the following sections.

6.1 System

For each age and speaker combination, we automatically extracted features from the speech signal, and combined them to create an unlabeled feature vector (UFV). Since we wanted to create a general age/speaker model, we did not consider the individual feature values or variations between frames of speech.

6.1.1 Gaussian Mixture Model

We used unsupervised clustering to first create a Universal Background Model. Starting off with our unlabeled feature vectors and 128 random points in the feature space represented as Gaussian functions, we conducted twenty iterations of the Expectation Maximization (EM) algorithm to create a Universal Background Model (UBM). We decided 128 points would be sufficient enough for our model because this number would be able to model the 60 or so phonemes that would describe all of the world's languages, and would also have enough tolerance to model many of the common bigrams and trigrams. Using the UBM as our starting point, we conducted Maximum A Posteriori (MAP) adaptation with one iteration of the EM algorithm to create an age/speaker adapted model for each age and speaker.

6.1.2 Support Vector Regression (SVR)

We tested using by using Support Vector Regression from Weka (Witten and Frank, 2005). Similar to our training set, we created a testing set by extracting the features from speech of a target age/speaker to create an unlabeled feature vector. With our testing set feature

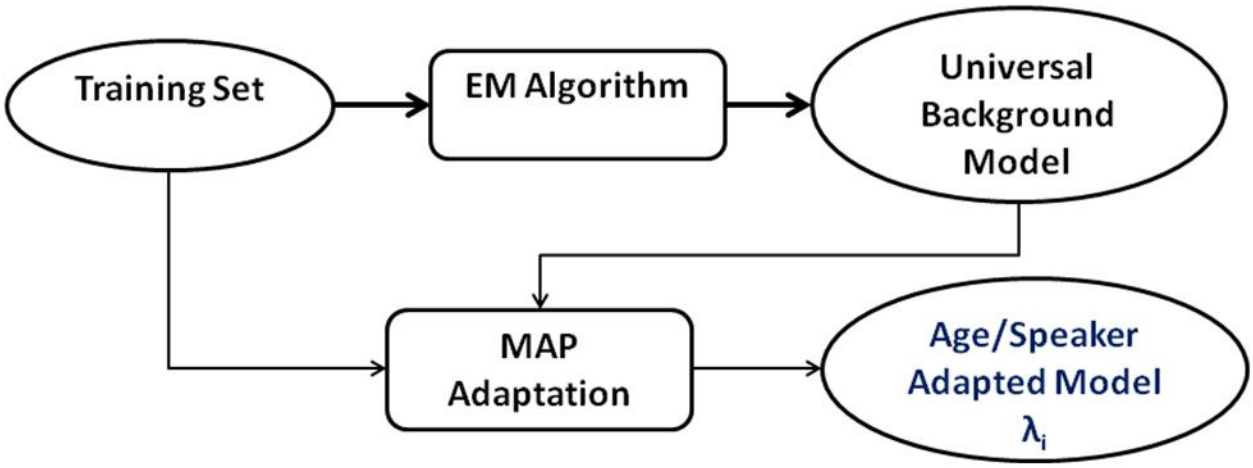

Figure 6.1: Age/speaker model creation process of our system using UBM/GMM

vectors and our UBM from training, we adapted our features using MAP estimation create a model of the test speaker.

Using our age/speaker models from both testing and training, we conducted SVR on the test data to predict the age of the target speaker. We used four different kernels for evaluation, namely linear, third degree polynomial, radial basis function with delta 0.1 and 0.01.

6.2 Results

For longitudinal data (Queen and Cooke broadcast data), we evaluated using a 30-fold leave one age out cross validation on the training set. In other words, we trained on 29 of the 30 recordings given at different ages, and tested on the recording for the one remaining age. We repeated this process 30 times for each age. For the UF-VAD data, we split the data into three sections, trained on two sections, and tested on the remaining section. We repeated this process three times each for a different testing set, and averaged the output to get the final result.

6.2.1 Regression Results

The following tables show the Pearson correlation and mean absolute error of evaluation results. Each table shows the correlation and error for the MFCC, formant frequency, and Maeda parameter features using the linear (Linear), third-degree polynomial (Poly3), radial basis function with delta 0.1 (RBF0) and 0.01 (RBF1). Best performance for each dataset is in bold. All regression results showed best performance with MFCC features using a linear kernel for regression. The high correlation rate in UF-VAD for males and females show that our features show a high correlation with age regardless of channel conditions. Also, even though formant frequencies have only three dimensions compared with the 12 dimensions for the MFCCs, we were able to get considerably high correlation rates. This shows that a

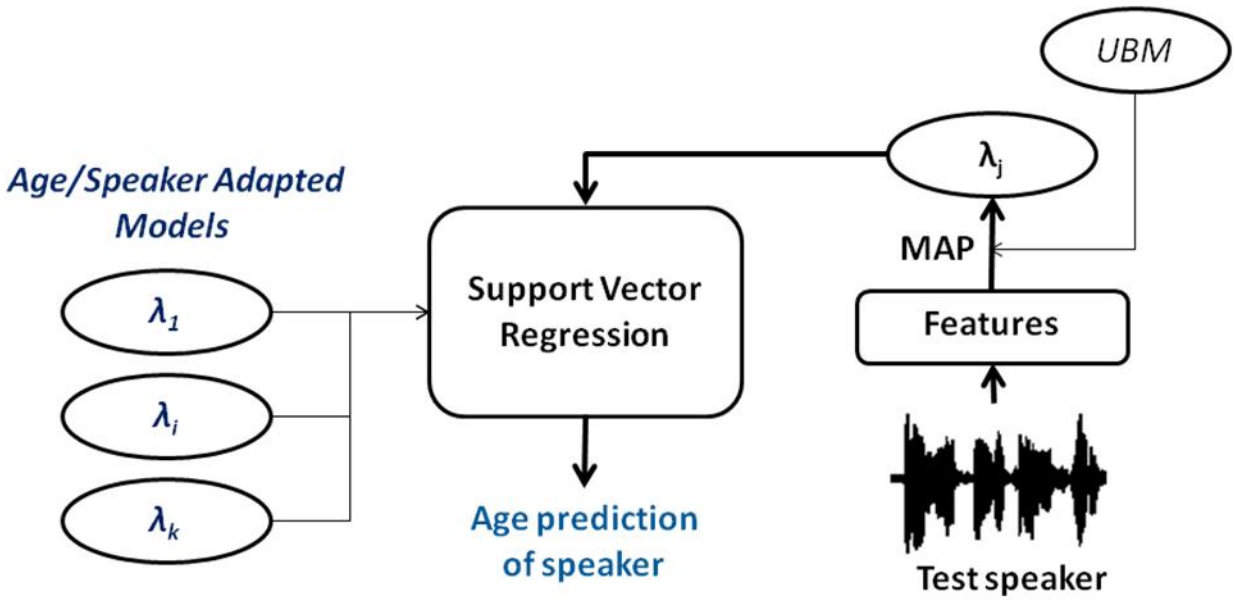

Figure 6.2: Age prediction process of our system using support vector regression (SVR)

lot of age information is also included in formant frequencies.

6.2.2 Speech Lengths for Accurate Age Classification

We also looked at the length of speech recordings needed to create an accurate model of the age and speaker by comparing regression results with different lengths of recordings from the UF-VAD. We were able to see that performance reaches its maximum at around 60 seconds as shown in the figure below, and were able to conclude that speech lengths at around one minute contains enough information to conduct accurate age classification.

6.2.3 Age Prediction: Human vs. System

We compared our classification results of the UF-VAD with classification results of an average of 40 humans, and were able to show that our system did not perform much differently from that of humans.

Human Performance

Forty people were asked to classify the age of the speaker for each of the 150 speakers in the database (Harnsberger, 2008). The average estimated age of the forty people were used for the following evaluation. The confusion matrix comparing the actual age and the predicted age is shown in the table 6.5. The performance of an average of forty people had an accuracy of 87 % with a mean absolute error of 9 years.

		Pearson Correlation	Mean Absolute Error
MFCC	Linear	**0.95**	**4.2**
	Poly3	0.77	11.4
	RBF0	-0.43	16.0
	RBF1	0.87	11.4
Formants	Linear	0.92	5.3
	Poly3	0.84	8.4
	RBF0	0.84	8.5
	RBF1	0.93	5.3
Maeda	Linear	0.87	6.7
	Poly3	0.87	6.6
	RBF0	-0.31	16.0
	RBF1	0.90	6.6

Table 6.1: Age regression results: Queen of England

		Pearson Correlation	Mean Absolute Error
MFCC	Linear	**0.94**	**4.3**
	Poly3	0.67	14.1
	RBF0	-0.57	14.5
	RBF1	0.84	9.3
Formants	Linear	0.90	5.9
	Poly3	0.91	5.4
	RBF0	0.77	11.3
	RBF1	0.87	6.7

Table 6.2: Age regression results: Alistair Cooke

		Pearson Correlation	Mean Absolute Error
MFCC	Linear	**0.80**	**12.1**
	Poly3	0.77	19.1
	RBF0	N/A	20.9
	RBF1	0.67	18.7
Formants	Linear	0.31	22.1
	Poly3	0.52	17.2
	RBF0	0.50	18.7
	RBF1	0.66	15.3
Maeda	Linear	0.66	14.3
	Poly3	0.65	14.8
	RBF0	N/A	20.9
	RBF1	0.67	16.0

Table 6.3: Age regression results: UF-VAD Males

		Pearson Correlation	Mean Absolute Error
MFCC	Linear	**0.92**	**8.9**
	Poly3	0.89	18.8
	RBF0	N/A	20.6
	RBF1	0.83	16.5
Formants	Linear	0.72	13.7
	Poly3	0.78	11.6
	RBF0	0.77	16.8
	RBF1	0.78	12.0
Maeda	Linear	0.83	11.3
	Poly3	0.83	11.2
	RBF0	0.62	20.3
	RBF1	0.82	12.1

Table 6.4: Age regression results: UF-VAD Females

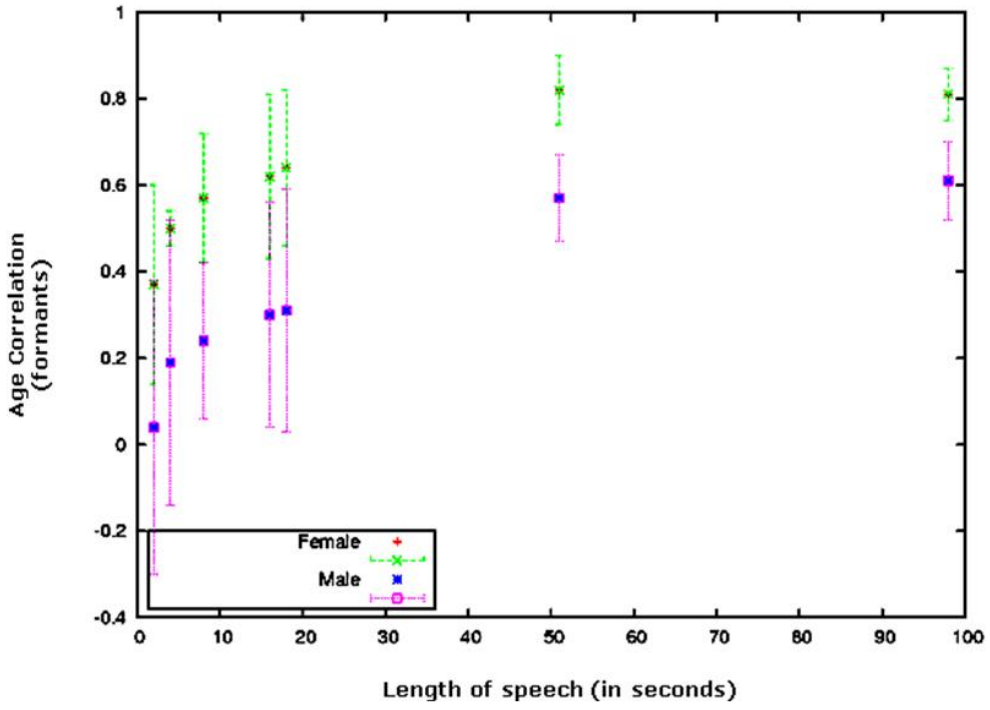

Figure 6.3: Age prediction process of our system using support vector regression (SVR)

		Actual		
	Age	35	35-59	59
Predicted	35	48	4	0
	35-59	2	46	14
	59	0	0	36

Table 6.5: Confusion matrix of human age classification

		Actual		
	Age	35	35-59	59
Predicted	35	29	1	0
	35-59	20	43	8
	59	1	6	42

Table 6.6: Confusion matrix of system age classification

System Performance

We conducted the same evaluation on our regression system. The confusion matrix is shown in table 6.6. Our system had an accuracy of 76 % with a mean absolute error of 11 years, which is not much different from human results.

6.3 Conclusion

We examined various features that correlated with age, and created a system that would estimate a person's age using speech. Using this system we evaluated MFCC, formant frequency and Maeda features and were able to show that these features had a high correlation with age. We also compared the performance of our system with age prediction evaluations of 40 humans and were able to show that there was not much difference between the performance of our system and humans.

Chapter 7

Summary

We report on investigations, conducted at the 2008 JHU Summer Workshop, of trying to predict the age of speaker based on the acoustic signal. Besides standard features used in automatic speech processing (MFCCs) we looked at a large vector of prosodic features and implemented a glottal excitation model and an articulatory inversion model to derive articulatory and excitation parameters. All these parameters (or there changes over time) were examined as indicators for changes in the acoustic signal due to aging processes. To do this we looked at dataset which were collected over more than 50 year from two different speakers (The Queen of England and Alistair Cooke) in relatively constant communication settings. In order to verify our findings and to exclude systematic changes based on changing recording conditions rather than aging voices or aging processes we also looked at many speakers of differen age groupes who were recorded under identical conditions (UF-VAD).

Our implementation of the glottal excitation model is based on eight parameters. Using short time frames (25ms), we minimized the distance between the synthesized excitation signal and the LPC residue to find optimal model parameters. We analyzed the correlation of the model parameters with age on longitudinal and cross-sectional data. The experiments provided evidence that

1. age information is contained in the glottal excitation signal,

2. this information may be extracted from a reduced physical model of the glottis with few parameters.

The implementation of Maeda's articulatory model has proven its effectiveness for acoustic-to-articulatory inversion. Correlating the articulatory parameters to age, one of the most striking observation on the trend was the increasing value of the average of P1, which indicates the position of the jaw. The tongue shape and the tongue tip parameters (P3 - P4) were with P7 the only parameters that did not see their standard deviation decrease with age. The model adaptation performed on the Queen indicates an increase of the length of the pharyngeal cavity with age, which is consistent with what we expected from the literature. This lengthening of the pharynx was however quite slow, and since this study was only performed on one speaker, these results need to be taken with caution.

Predicting the age of the speakers we examined various features that correlated with age, and created a system that would estimate a person's age using speech. Using this system we

evaluated MFCC, formant frequency and Maeda features and were able to show that these features had a high correlation with age. We also compared the performance of our system with age prediction evaluations of 40 humans and were able to show that there was not much difference between the performance of our system and humans.

Bibliography

[Ahme 74] N. Ahmed, T. Natarajan, and K. Rao. "Discrete Cosine Transform". *IEEE Transactions on Computers*, Vol. 23, pp. 90–93, 1974.

[Batl 00a] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke. "The Prosody Module". In: W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translations*, pp. 106–121, Springer, New York, Berlin, 2000.

[Batl 00b] A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer. "The Recognition of Emotion". In: W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translations*, pp. 122–130, Springer, New York, Berlin, 2000.

[Bera 49] L. L. Beranek. *Acoustic Measurements*. Wiley, New York, 1949.

[Buck 99] J. Buckow, V. Warnke, R. Huber, A. Batliner, E. Nöth, and H. Niemann. "Fast and Robust Features for Prosodic Classification". In: V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, Eds., *Text, Speech and Dialogue, 2nd International Workshop, September 13-17, 1999, Plzen, Czech Republic, Proceedings*, pp. 193–198, Berlin, 1999.

[Davi 80] S. B. Davis and P. Mermelstein. "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366, 1980.

[Durb 60] J. Durbin. "The Fitting of Time Series Models". *Review of the International Statistical Institute*, Vol. 28, No. 3, pp. 233–244, 1960.

[Fant 60] G. Fant. *The Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[Fant 68] G. Fant. "Analysis and synthesis of speech processes". In: B. Malmberg, Ed., *Manual of Phonetics*, Chap. 8, pp. 173–276, North-Holland Publ. Co, Amsterdam, 1968.

[Ferr 02] C. Ferrand. "Harmonics-to-Noise Ratio: An Index of Vocal Aging". *The Journal of Voice*, Vol. 16, No. 4, pp. 480–487, 2002.

[Furu 86] S. Furui. "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 34, No. 1, pp. 52–59, 1986.

[Gall 02] F. Gallwitz, H. Niemann, E. Nöth, and V. Warnke. "Integrated Recognition of Words and Phrase Boundaries". *Speech Communication*, Vol. 36, No. 1-2, pp. 81–95, 2002.

[Galv 97a] A. Galván-Rodrigez. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives.* Thèse de l'Institut National Polytechnique de Grenoble, 1997.

[Galv 97b] A. Galván-Rodrigez. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives.* Thèse de l'Institut National Polytechnique de Grenoble, 1997.

[Harn 08] J. D. Harnsberger, R. Shrivastav, W. S. Brown, H. Rothman, and H. Hollien. "Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age". *The Journal of Voice*, Vol. 22, No. 1, pp. 58–69, 2008.

[Harr 06] J. Harrington. "An acoustic analysis of 'happ-tensing' in the Queens's Christmas broadcasts". *Journal of Phonetics*, Vol. 34, pp. 439–457, 2006.

[Hube 02] R. Huber. *Prosodisch-linguistische Klassifikation von Emotion. Studien zur Mustererkennung*, Logos Verlag, Berlin, 2002.

[Kies 97] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik*, Shaker, Aachen, 1997.

[Kirk 83] S. Kirkpatrick, C. Gelatt, and M. Vecchi. "Optimization by Simulated Annealing". *Science*, Vol. 220, No. 4598, pp. 671–680, 1983.

[Komp 97] R. Kompe. *Prosody in Speech Understanding Systems.* Vol. 1307 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin, 1997.

[Lapr 98] Y. Laprie and B. Mathieu. "A variational approach for estimating vocal tract shapes from the speech signal". In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 929–932, Seattle, USA, May 1998.

[Leeu 04] I. Verdonck-de Leeuw and H. Mahieu. "Vocal aging and the impact on daily life: a longitudinal study". *The Journal of Voice*, Vol. 18, No. 2, pp. 193–202, 2004.

[Levi 47] N. Levinson. "The Wiener RMS Error Criterion in Filter Design and Prediction". *Journal of Mathematics and Physics*, Vol. 25, pp. 261–278, 1947.

[Linv 02] S. E. Linville. "Source Characteristics of Aged Voice Assessed from Long-Term Average Spectra". *The Journal of Voice*, Vol. 16, No. 4, pp. 472–479, 2002.

[Linv 96] S. E. Linville. "The Sound of Senescence". *The Journal of Voice*, Vol. 10, No. 2, pp. 190–200, 1996.

[Maed 79] S. Maeda. “Un modèle articulatoire de la langue avec des composantes linéaires”. In: *Actes 10èmes Journées d'Etude sur la Parole*, pp. 152–162, Grenoble, Mai 1979.

[Maed 90] S. Maeda. “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model”. In: W. Hardcastle and A. Marchal, Eds., *Speech production and speech modelling*, pp. 131–149, Kluwer Academic Publisher, Amsterdam, 1990.

[Nait 99] M. Naito, L. Deng, and Y. Sagisaka. “Model-based speaker normalization methods for speech recognition”. In: , Budapest, September 1999.

[Neld 65] J. A. Nelder and R. Mead. “A Simplex Method for Function Minimization”. *The Computer Journal*, Vol. 7, No. 4, pp. 308–313, 1965.

[Noth 00] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. “Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System”. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 5, pp. 519–532, 2000.

[Noth 02] E. Nöth, A. Batliner, V. Warnke, J.-P. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. “On the Use of Prosody in Automatic Dialogue Understanding”. *Speech Communication*, Vol. 36, No. 1-2, pp. 45–62, 2002.

[Noth 88] E. Nöth and R. Kompe. “Der Einsatz prosodischer Information im Spracherkennungssystem EVAR”. In: H. Bunke, O. Kübler, and P. Stucki, Eds., *Mustererkennung 1988 (10. DAGM Symposium)*, pp. 2–9, Berlin, 1988.

[Noth 91] E. Nöth. *Prosodische Information in der automatischen Spracherkennung – Berechnung und Anwendung*. Niemeyer, Tübingen, 1991.

[Olss 75] D. Olsson and L. Nelson. “The Nelder-Mead simplex procedure for function minimization”. *Technometrics*, Vol. 17, No. 1, pp. 45–51, 1975.

[Over 62] J. E. Overall. “Orthogonal factors and uncorrelated factor scores”. *Psychological Reports*, pp. 651–662, 1962.

[Pota 07] B. Potard and Y. Laprie. “Compact representations of the articulatory-to-acoustic mapping”. In: *Interspeech, Anvers*, Aug. 2007.

[Riec 95] S. Rieck. *Parametrisierung und Klassifikation gesprochener Sprache*. Vol. 353 of *VDI Fortschrittberichte Reihe 10: Informatik/Kommunikationstechnik*, VDI Verlag, Düsseldorf, 1995.

[Scho 97] J. Schoentgen and S. Ciocea. “Kinematic formant-to-area mapping”. Vol. 21, pp. 227–244, 1997.

[Schu 95] E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen*. Vieweg Verlag, Braunschweig, Wiesbaden, 1995. freely available at `http://www.minet.uni-jena.de/fakultaet/schukat/asebuch.html`, last visited 01/18/2008.

[Soro 00] V. Sorokin, A. Leonov, and A. Trushkin. "Estimation of stability and accuracy of inverse problem solution for the vocal tract". Vol. 30, pp. 55–74, 2000.

[Stei 08] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Childrens's Speech*. PhD thesis, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 2008.

[Stem 05] G. Stemmer. *Modeling Variability in Speech Recognition*. Logos Verlag, Berlin, 2005.

[Stev 37] S. S. Stevens, J. Volkmann, and E. B. Newman. "A Scale for the Measurement of the Psychological Magnitude Pitch". *Journal of the Acoustical Society of America*, Vol. 8, No. 3, pp. 185–190, 1937.

[Stev 98] K. N. Stevens. *Acoustic Phonetics*. The MIT Press, Cambridge, MA 02141, 1998.

[Wahl 00] W. Wahlster, Ed. *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, New York, Berlin, 2000.

[Warn 03] V. Warnke. *Integrierte Segmentierung und Klassifikation von Äußerungen und Dialogakten mit heterogenen Wissensquellen*. Logos Verlag, Berlin, 2003.

[Wigh 92] C. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University Graduate School, 1992.

[Zeis 06] V. Zeißler, J. Adelhardt, A. Batliner, C. Frank, E. Nöth, R. P. Shi, and H. Niemann. "The Prosody Module". In: W. Wahlster, Ed., *SmartKom: Foundations of Multimodal Dialogue Systems*, pp. 139–152, Springer, Berlin, Heidelberg, 2006.

List of Figures

List of Tables