Fundamental Frequency Estimation and Modeling for Speaker Recognition

Rosa Emilia González Hautamäki

15.7.2005
University of Joensuu
Department of Computer Science
Master's thesis.

Abstract

Fundamental frequency (F_0) is the rate of vocal folds vibration during speech. It is considered to be one of the most important prosodic features to characterize speech and speaker specific patterns. This study considers two aspects: the estimation of F_0 and its modeling for speaker recognition. To understand this prosodic parameter, we present an analysis of three popular algorithms: autocorrelation function (ACF), average magnitude difference function (AMDF) and cepstrum analysis. Many attempts have been made to achieve accuracy in F_0 estimation.

F_0 has some advantages over spectral features, and it has been recently used for improving performance of speaker recognition systems. We present a model for long-term F_0, based in both parametric and non-parametric approaches. The first one refers to the statistical properties of the F_0 distribution, while the non-parametric approach analyze the density function of the distribution using histograms. These approaches are combined to describe the F_0 model. We present a method that combines a spectral feature, mel-frequency cepstral coefficients (MFCC) with F_0 for closed-set text independent speaker identification, and evaluate the methods using telephone-quality corpus of 180 speakers. The F_0 feature yields rather high error rate (89.4 %), under both matched and mismatched conditions. In the other hand, MFCC performs efficiently when the training and testing conditions are clean speech, but decreases its performance in the mismatched and noise cases. This study shows that F_0 has potential information to add to the recognition task.

Keywords: Text-independent speaker recognition, speech signal processing, fundamental frequency estimation, fundamental frequency modeling, classifier fusion. pitch determination algorithm.

Computing Reviews (1998): Signal processing (I.5.4), Signal analysis, synthesis, and processing (H.5.5), Natural language processing (I.2.7), Pattern recognition (I.5)

Acknowledgements

During the course of this research, I was fortunate to interact with a number of great people that influence the direction and quality of my work. First I would like to express my gratitude to Tomi Kinnunen for his outstanding guidance, support and cooperation. Tomi was always available to advise me on technical and theoretical problems and gave me the opportunity to learn about the fascinating world of signal processing and applications to speaker recognition. He encourage me to explore new ideas and try them. This was very difficult and I really appreciate his patience to supervise my work, I learn from the hard way and from the mistakes too, but would not have been so joyful without a diligent supervisor like him.

The Department of Computer Science in Joensuu gave me the opportunity to be part of the IMPIT program and I am grateful for the learning experience I got from being expose to such a learning environment, I thank all my teachers and assistants for open my mind to new knowledge.

My work in this remote place would not be possible without the constant inspiration, support and love of my family in El Salvador. I dedicate my work to my parents, Dora and Victor for being the inspiration of my life; to my sisters Cristina and Séfora and my brother Efraín, for having the right words that encourage me to do the best effort. Thanks to the rest of my family and friends who have been attentive to my progress in this country. Also great thanks goes to my family in Finland, without them life in this country would have been really difficult.

Thanks to my friends in Joensuu, specially the bahá'í friends for their love and care that help me to enjoy my new life in this city. These first years have been happy thanks to you.

These would not be complete without the thanks to my loving husband Ville, for all the support, encouragement, help and caring. Certainly this work would not be possible without a colleague like him. I am fortunate to have you the rest of my life.

Contents

1 Introduction **1**

1.1 Automatic speaker recognition 2

1.2 Fundamental frequency as a feature 3

1.3 Purpose and contents of this study 4

2 Speech Production **5**

2.1 Description of voice production 5

2.2 Speech acoustics 11

3 Fundamental Frequency Estimation **15**

3.1 Problem definition 15

3.2 Fundamental frequency estimation 16

3.2.1 Categorization of Pitch Determination Algorithms (PDA) . . . 17

3.3 Short-term analysis pitch determination 18

4 Short-term fundamental frequency estimation **21**

4.1 Autocorrelation function 21

4.2 Average Magnitude Difference Function 24

4.3 Cepstrum analysis 27

5 Fundamental frequency for speaker recognition **31**

5.1 Speaker recognition as a pattern recognition problem 31

5.2 Modeling the F_0 distribution 32

5.3 Parametric model 33

5.4 Non-parametric model 36

5.5 Mel-Frequency Cepstral Coefficients 37

5.6 Classifier fusion 38

6 Experiments **40**
6.1 Experimental setup 40
6.2 Speaker Modeling 42
6.2.1 Non-Parametric F_0 model 43
6.2.2 Parametric F_0 model 44
6.2.3 MFCC model 44
6.3 Results 45
6.3.1 Results for F_0 classifiers 45
6.3.2 Combining F_0 and MFCC 47
6.3.3 Samples correctly classified by F_0 but not by MFCC 48
6.4 Summary of results 50

7 Conclusions **51**
7.1 F_0 extraction 51
7.2 F_0 modeling for speaker recognition 52

A Tuning set **58**

Chapter 1

Introduction

The identification of people by their voices is a common practice in everyday life. We identify persons by listening to their voices, over a phone line, radio, among other devices. If the person is familiar to us, we can identify her/him by the tone of the voice, the style of speaking, and so on. If we do not know her/him, we can still infer some characteristics like gender, age, emotional state and language, among others.

Speech is produced by the vibration of the vocal folds modified by the resonance of the vocal tract [17, 38, 21, 5]. It expresses many levels of information, including semantic, linguistic, articulatory and acoustic one [5]. The aim of *speaker recognition* systems is to extract, characterize and recognize this information from the speech signal [37]. Speaker recognition is a *performance biometric*, which means that the person has to perform a task to be recognized: give a sample of his/her voice.

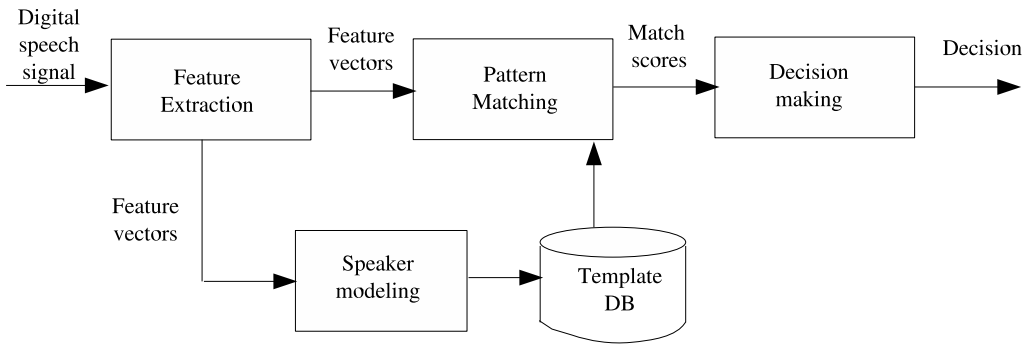


Figure 1.1: General scheme for speaker recognition

1.1 Automatic speaker recognition

Traditional pattern recognition models, such as, speaker recognition, consist of three components (See Fig. 1.1): *feature extraction* and *selection*, *pattern matching* and *classification*. These phases are interrelated and their performance is not independent. The aim is to design a system that minimizes the recognition error and as a result, to differentiates between the speakers.

The selection of good features is of main importance. *Feature extraction* is the estimation of variables (feature vector) from the observation of a speech signal. The idea is to transform the data contained in the signal into a collection of variables that can preserve the information and that can be used to make comparisons.

In [38], the author list the following properties for ideal features:

- high inter-speaker variation,
- low intra-speaker variation,
- easy to measure,
- robust against disguise and mimicry,
- robust against distortion and noise,
- maximally independent of other features.

In the training phase, the features are used to create the speaker model that will be used to match the correct speaker in the recognition phase.

Speaker recognition is divided in two main tasks [5, 36, 37]: *identification* and *verification*. The identification task is to recognize who is talking, from a set of previously stored voices. If the unknown voice comes from a set of known speakers, the task is called *closed-set* identification. In verification, the task is to decide if the person is who claims to be (yes/no answer). In general an impostor, will be classified as not known by the system and refers as "unknown", then the set is called *open-set*.

Speaker recognition systems can be further divided into *text-dependent* or *text-independent* systems. In the first case, the system has prior knowledge of the content of the utterance, while in the second case, there is no prior knowledge of the text to be uttered. The classifier uses the features extracted, either to identify or verify the speaker. During the recognition phase, the classifier associates a similarity or distance measurement for the unknown speaker vectors in reference to the stored models. The model that obtains the best score gives the identity of the speaker.

1.2 Fundamental frequency as a feature

Many features of speech are used in speaker recognition systems. The goal is to find a feature or group of features that can successfully discriminate one speaker from another one. Two main sources of speaker characteristics are physical and learned [5]. The vocal tract shape, larynx, vocal folds and glottis are important organs involved in the speech production and are part of the physical factor of speech. Some learned characteristics of speech production include speaking rate and dialect.

Fundamental frequency (F_0) is the rate of the vocal folds vibration, and it is said to be useful for speaker discrimination [38, 6, 11]. In speaker recognition systems, the emphasis of this feature has been in the forensic field [38, 31], and has been proved to be useful prosodic parameter for identification of speakers as it contains good speaker specificity not only between male and female speakers, but also between speakers of the same sex that have lower or higher average F_0 values [38, 1, 11]. The F_0 can also help to distinguish lexical categories in tonal languages, word segmentation, gender separation, among other uses [1, 11, 6, 20]. In general, F_0 estimation is an useful element in many signal processing methods, e.g. as a metadata for multimedia content indexing [9]. There are factors that can make the glottal vibrations aperiodic, such as, glottalizations, *vocal creak* or *fry*, easily impersonation, changes in the amplitude, and so on [9]. These factors difficult the task of obtaining a reliable estimation of F_0, which still an open problem. However, F_0 is more independent of communication channel than spectral coefficients as it is not or only slightly affected by the noise as well as by the channel specificities [3, 18]. For this reason, there has been attempts to fuse F_0, as a prosodic feature, and spectral features to improve speaker recognition [19, 11, 43, 22].

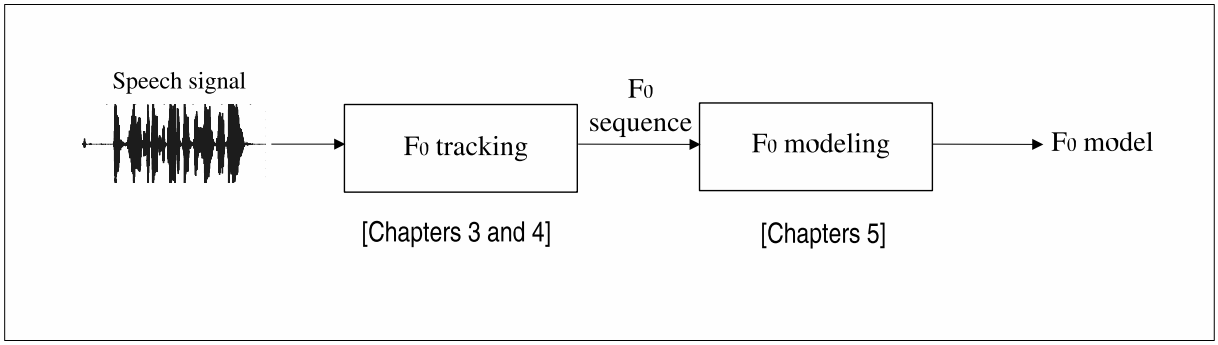


Figure 1.2: Components of proposed method.

1.3 Purpose and contents of this study

Fundamental frequency is the most studied prosodic parameter that characterizes speech. The goal is to analyze the influence of F_0 in speaker recognition and find effective methods to estimate and use it. In particular, we aim to incorporate F_0 as a feature for speaker recognition in combination with spectral features. We consider the F_0 distribution with its statistical properties and its density function represented by histogram, as a speaker model. This model is further fused with a spectral feature, MFCC, to perform closed-set text-independent speaker recognition. The study of this method is done under matched and mismatched conditions.

The rest of the thesis is organized as follows: In chapter 2, we present an overview in the speech production process from the articulatory point of view. Chapter 3 covers the main issues of F_0 estimation, voicing determination and categorization of F_0 detection algorithms. Chapter 4 outlines three short-term fundamental frequency estimation algorithms. Chapter 5 presents a model to incorporate F_0 in speaker recognition systems. Chapter 6 describes the experimental setup and results of this study and Chapter 7 concludes with a summary of the analysis of this work.

Chapter 2

Speech Production

The physical production of speech has been explained most frequently with the aerodynamic theory. In this chapter we present a brief description of this process from the articulatory and acoustic points of view.

2.1 Description of voice production

To understand the speech production process it is necessary to mention the anatomy related with it and the functions that different organs have. The field of *articulatory phonetics* studies how speech sounds are produced and how the structures of the vocal tract interact. As a brief summary to speech production, we concentrate on the functions of the larynx and the articulatory organs, that include oral and pharyngeal cavities (see Fig. 2.1).

The main physiological aspect of the human speech production system is at the vocal tract. The vocal tract consists of the following parts: (1) *laryngeal pharynx* (under the epiglottis), (2) *oral pharynx* (between the epiglottis and velum), (3) *oral cavity* (limited by the lips, tongue, and palate), (4) *nasal pharynx* (above the velum, end of nasal cavity), and (5) *nasal cavity*.

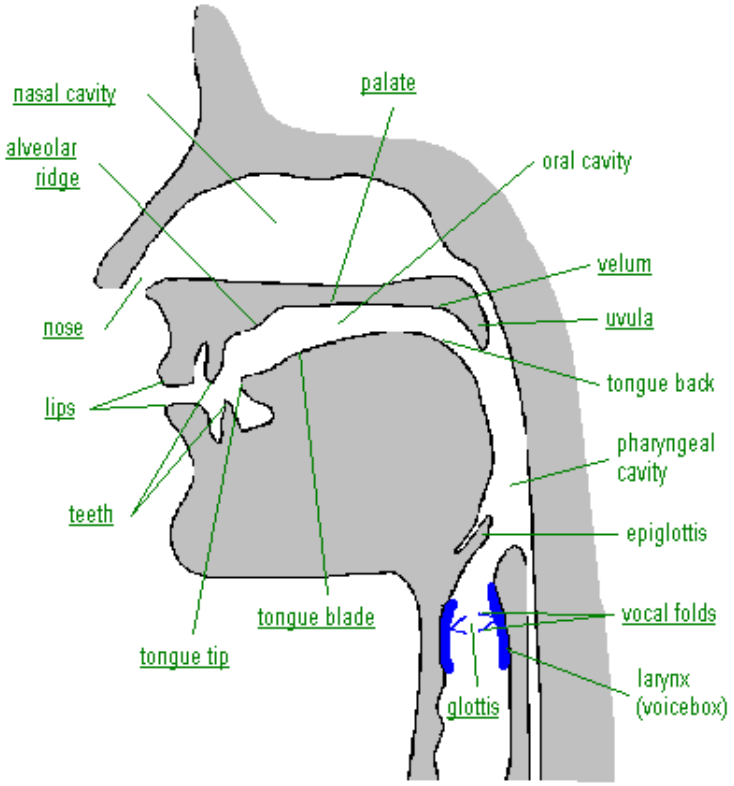


Figure 2.1: The principal organs for articulation [33].

The *larynx* or *vocal box* (see Fig. 2.2), has the functions of swallowing, breathing and phonation (voice production). Voice production can be seen as a result of the production of airflow, resonance sound and articulation of voice. These aspects are described in the following sections.

Phonation mechanism

According to Rose [38], vocal cords and supralaryngeal vocal tract are basic structures for the production of speech, and represent two independently functioning and controlled modules.

The lungs provide with the necessary airflow to overcome the tension of closed *vocal folds*. Vocal folds (or vocal cords) are elastic ligaments attached inside the walls of the larynx (see Fig. 2.2), and they can be manipulated to be open (*abducted*) or closed (*adducted*), and be *tensed* or *relaxed*. The space between the vocal folds is called *glottis* and its function is to let the airflow to pass through the trachea, or if it is closed, to stop the air stream.

The vocal folds vibration is called *voicing* or *phonation*. Voiced sounds are produced at the larynx as a repetition of events. First, the vocal folds are adducted, blocking the flow of air from the lungs. The subglottal pressure increases until the resistance of the vocal folds is overcome, and they open again. Then they close rapidly by a combination of factors like their elasticity, tension of the laryngeal muscle and *Bernoulli effect* [40, 21]. The process continues and if it is maintained with a steady supply of pressured air, the vocal folds will open and close in a periodic way. The sound produced by the larynx travels through the throat and mouth where, it is later modified to produce speech.

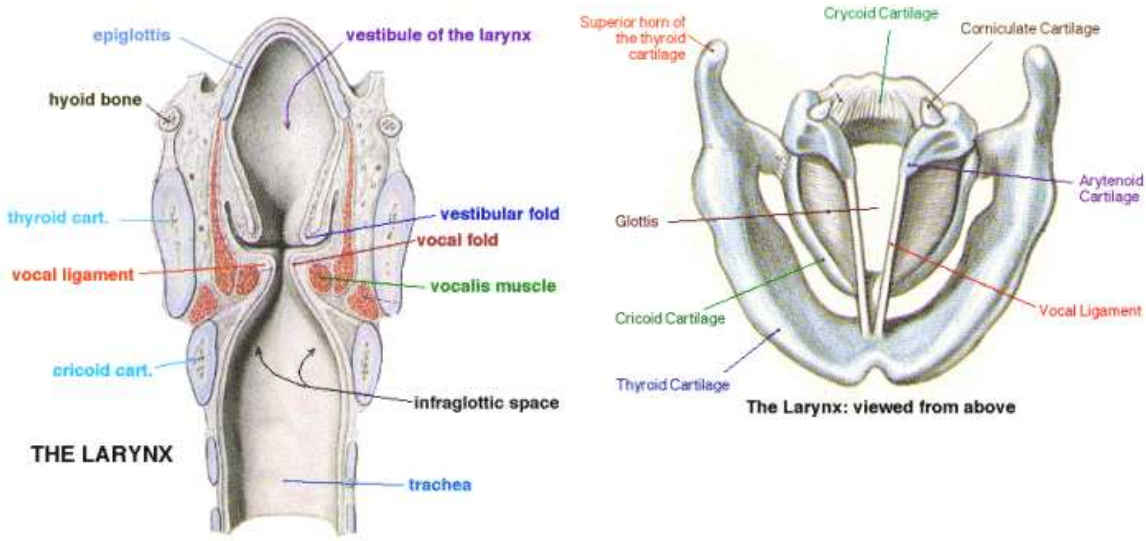


Figure 2.2: Diagrams with views of the larynx [14].

Figure 2.3 shows a simplified diagram of the vocal folds and sound production. First, vocal folds are together (1), air is forced to the trachea pushing them until the upper edge of them opens (2, 3, 4). Then the air passes through the glottis with an increased velocity, which implies a pressure drop at the glottis. As a result, the vocal folds start to adduct again, starting from the lower edge (6-8). The frequency of oscillation is called the *fundamental frequency*, and it is a characteristic physically based in the length, tension and mass of the vocal folds [5, 38].

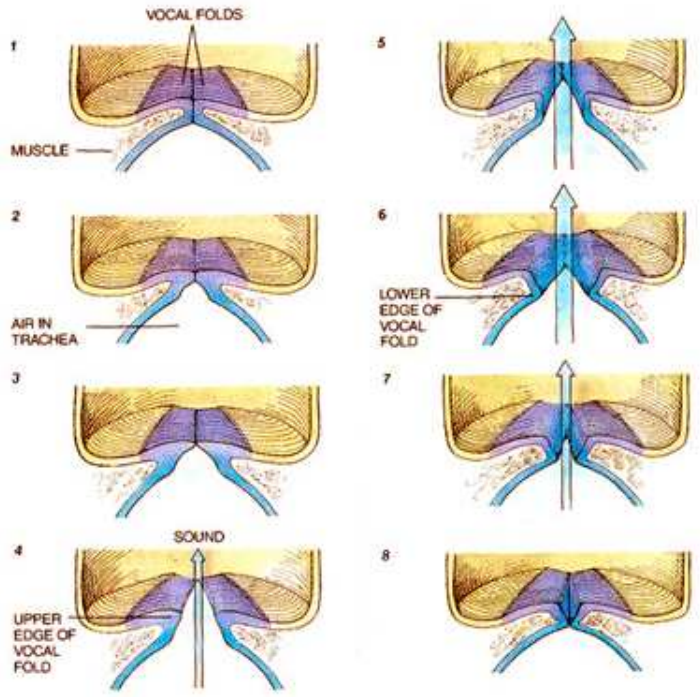


Figure 2.3: Diagram of vocal folds in the production of sound [25].

Speech sounds can be classified according to the phonation type as *voiced*, *unvoiced* and *whispered sounds*. Voiced sounds are produced by the vibration of the vocal folds. All vowels and certain consonants like /m/, /n/, /l/, /r/, /z/ are voiced sounds [38]. Unvoiced (voiceless) sounds lack the vibration of the vocal folds and their production is characterized by airstream through the open glottis. Examples are the consonant /h/ in Finnish language, as in the first letter in the word "*hattu*"('*hat*') and /j/ in Spanish like in the word "*jamón*"('*ham*'). Whispered sounds are produced by airflow through a small opening between the arytenoid cartilage (see Fig. 2.2) at end of almost closed vocal folds. According to Hess [17], *unvoiced* sound segments can be of two types: *voiceless*, if there is a turbulence in the vocal tract producing a noiselike signal and there is not vibration of the vocal folds the segment, and *silence* that is a segment where there is no vocal source activity.

Supralaryngeal vocal tract

Many sounds are produced with the organs above the vocal folds to the lips, that is a module named *supralaryngeal vocal tract* and it includes oral cavity, pharynx and nasal cavities. The vocal tract is responsible for the resonance effect for the produc-

tion of vowels and consonants, where the shape adopted by the cavities influence the sound. For example, *fricatives* are produced by constrictions in the vocal tract, like the consonants /s/ and /z/ in the words *sign* and *zoo*, respectively. In these sounds, the front part of the tongue moves to create a narrow constriction, in which the air becomes turbulent as it passes, creating an acoustic noise.

Nasal cavity also works as a resonance component. Air flows through the nasal cavity when the soft palate (*velum*) is down, that causes the resonance. If the soft palate is up, it closes the nasal cavity and there is no nasal resonance. In English, there are three important nasal sounds: /m/, /n/ and /ng/, like in the words *simmer*, *sinner* and *singer*, respectively.

Voicing and pitch

The voicing process is mostly a contribution from the opening and closing of the vocal folds, and the frequency of this pattern is known as fundamental frequency. Fundamental frequency is called also F_0 and *pitch*. The inverse of the F_0 is defined as *fundamental period* and the opening-closing cycle related to it is called *glottal pulse* (see Fig. 2.4).

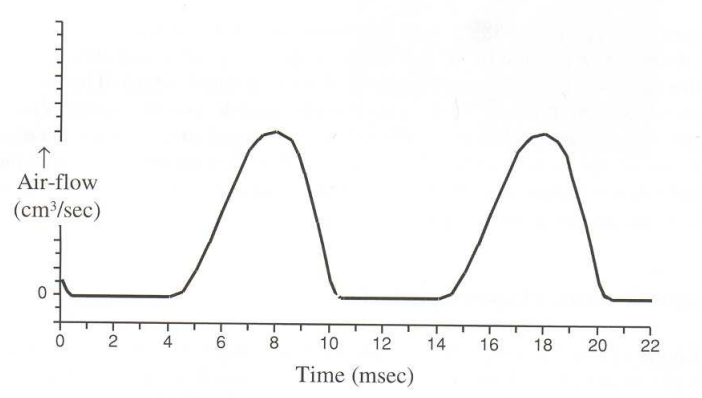


Figure 2.4: Two periods waveform of a glottal pulse [38]

The vocal folds are closed during the *close phase* stopping the air to flow through the glottis, as 0 to 4 ms in the first period. The *opening phase* refers to the process of the vocal folds come apart allowing the air through the glottis; the air increases until maximum then decreases and the vocal folds come together again which correspond to the 5 to 9 ms. The rate of decreasing is most rapid, as the closing phase is shown in the diagram from 9.5 ms to 10.5 ms. Vocal folds activity is an important part of

the study of the fundamental frequency, as it can differ from speech sounds in three ways: (1) absence or presence of vibration, (2) differences in rate vibration or (3) by differences in mode of vibration [38].

For voicing, the key is to determine between voiced and voiceless sounds that are linked to the tension of the vocal folds and the subglottal pressure. The frequency of the vibration rate can be controlled by changing the tension during the vibration phase. Rose [38] explains that increased tension produces higher frequencies while relaxation results in lower ones.

The fundamental frequency (F_0) has many uses in speech. Tonal languages make differences between simple sentences and questions or between phonemes in words that have similar pronunciation but different meanings. As an example, 'papa' and 'papá' in Spanish. The first one means *potato*, and is read by stressing the first syllable, the second word means *father* and it is read with emphasis in the second syllable.

There are studies that define differences between F_0 for male, female and children, due to the anatomic differences [21]. The average F_0 for European languages are approximately 120 Hz for males, 220 Hz for females and 330 Hz for children [21].

One determinant factor for the vibration of the vocal folds is its size, specifically mass and length. There are studies [38, 44] that compare the behavior of the vocal folds with a string and spring components. Since vocal folds stretch from the back of the larynx to the front, they are considered to have a *string-like* component. Each vocal fold behaves like a mass attach to a *spring* because of the medialateral movement. According to Titze [44], the F_0 of vocal folds can be described by Eq. (2.1) when they behave like a string.

$$F_0 = \frac{1}{2L_m}\sqrt{\frac{\sigma_c}{\rho}} \tag{2.1}$$

where L_m represents the length of the vocal folds in meters, σ_c is the longitudinal tension of the cords divided by the cross-sectional area of vibrating tissue, quantified in pascals (Pa); and ρ is the density of the cord.
The relationship between the F_0 and the length of the vocal folds is inversely proportional. For long vocal folds, the speaker has low F_0. Shorter vocal folds produce higher F_0.

The fundamental frequency of the focal folds when they behave like a spring is defined as [44]:

$$F_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{2.2}$$

where m represents the vibrating mass of the vocal folds in kilograms, and k is their stiffness quantified in newtons per meter. The F_0 of a great mass correspond lower values.

Then the F_0 can be related to the speaker anatomy and constitutes a within speaker characteristic.

2.2 Speech acoustics

Acoustic parameters, such as type of phonation and periodicity of the vocal cords wave, are considered important in the speech production process [17]. Harrington and Cassidy [15] mention four processes of the speech production, whose acoustic effects are considered independently. These are: (1) *sound source*, that can be by the vibration of the vocal folds (voiced sounds), or a turbulent airstream (voiceless sounds), or a combination of both (voiced fricatives), (2) *vocal tract filter*, which is the acoustic term for the shape of the vocal tract, (3) *energy losses* and, (4) *radiated sound pressure*, the way how a speech waveform radiates from the mouth.

Acoustic theory of speech production also known as *source-filter theory* explains the radiated acoustics in terms of vocal mechanism [12, 15].

The idea of this model is to explain the almost independent contribution of the *source* (vocal fold vibration) and the *filter* (modification of parameters like: shape, rate, energy, etc.) in the production of speech[15].

The source of acoustic energy is at the larynx during production of speech. The opening-closing cycle of vocal folds is repeated as long as phonation activity is taking place. One representation of the acoustic characteristics for the vocal folds vibration cycles is the glottal waveform, that it is a plot of the volume of airflow through the glottis by time as shown in Figure 2.4. It consist of three phases: (1) close phase, (2) opening phase and (3) closing phase. Then a fundamental period extends from the beginning of the opening phase to the end of the close phase.

The *spectrum* of the glottal waveform (see Fig. 2.5) consist of *amplitude* and *phase spectrum*. The amplitude spectrum displays amplitude as a function of frequency. The diagram shows that energy is present in the volume velocity wave with many frequencies known as *harmonics*. Then if F_0 is 100 Hz, the rest of frequencies are multiples of this (second and fourth harmonics have labels in the figure).

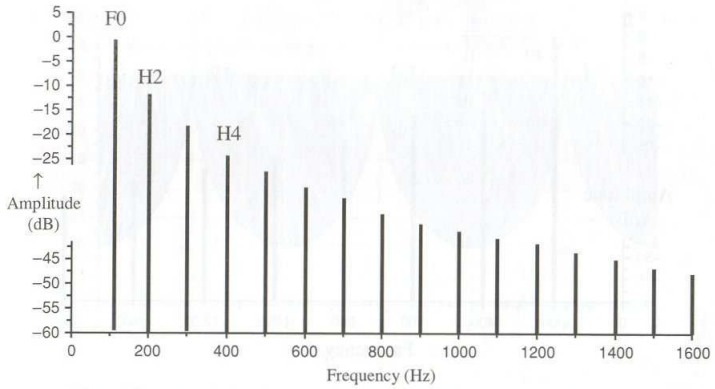


Figure 2.5: Spectrum of the glottal pulse waveform [38]

The more the speech wave differs from a sinusoidal shape, more harmonically sinusoidal components are needed to approximate it [15, 38].

A spectral characteristic of the glottal waveform is that the spectrum falls at a 12 dB rate every double of frequency, i.e. *12* dB/octave [15, 38]. For example, in Fig. 2.5, the fundamental frequency at 100 Hz has a difference of amplitude of 12dB with the double frequency at 200 Hz, and between this one and the 400 Hz there is also an amplitude difference of 12 dB. This difference in amplitude is known as *spectral slope*. The spectrum in the figure has a spectral slope of -12 dB/octave. The figures for the glottal pulse waveform and spectrum are related to voiced sounds and they can be used for characterizing the periodicity. Unvoiced and whispered sounds do not show a periodical waveform, and consequently the spectrum is non-harmonic. In this thesis, we concentrate on the voiced sounds as the main object is the estimation of the fundamental frequency directly related to these ones.

The task of the supralaryngeal vocal tract is to act as an acoustic filter that suppresses energy at some frequencies and amplifies others [38]. After the production of a sound, for example a vowel, the air coming from the lungs is interrupted by the vocal folds vibrations which allow a sequence of airstream to the supralaryngeal vocal tract. The airstream in the supralaryngeal vocal tract cause it to vibrate. The frequency and amplitude of the vibration, depends on the vocal tract shape. To explain that the shape of the supralaryngeal vocal tract , as a filter, modifies the source energy, there have been proposed model for its shape [15, 38].

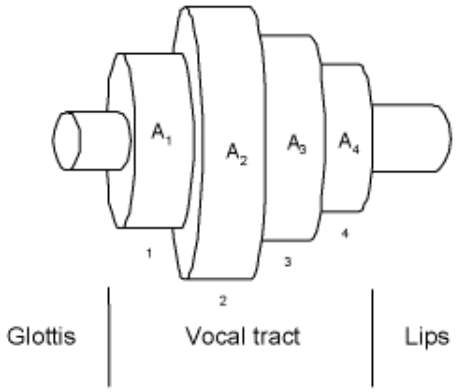


Figure 2.6: Acoustic tube model of speech production [21]

The vocal tract can be approximated as a set of interconnected cylinders, with a specific length referring to the vocal tract and an insignificant change for the diameters. The vocal tract is represented as an *area function* specified by the cross-sectional area and the length of each cylinder. The more cylinders are used, the more the model approximates the shape of the vocal tract. The way the air vibrates in the supralaryngeal vocal tract, given a particularly shape, can be represented by a frequency-amplitude spectrum called *transfer function* [17], from which the resonance of the interconnected cylinders can be calculated. The simplified model assumes no energy losses when the airstream passes through the interconnected cylinders (lossless tubes), due to a good approximation for the estimation of resonances of some sounds.

The frequencies at which there is a maximum energy in the spectrum of the transfer function are called *resonant frequencies*, and in acoustic phonetics the vocal tract resonances are known as *formants*.

Production of vowels involves less modifications of the vocal tract tube and approximates to the model with a uniform cross-sectional area, as when the tongue moves from front to back and high to low, changing the size of the mouth opening by spreading nor rounded.

The *schwa vowel* is the central vowel sound, typically occurring in weakly stressed syllables, as in the final syllable of 'sofa' and the first syllable of 'along' [10]. For the schwa vowel, the supralaryngeal vocal tract approximates a uniform tube, closed at one end (larynx). The resonant frequencies at which the air vibrates in a tube

of cross-sectional area closed at one end, can be described in function of the tube's length and can be calculated with the Eq. (2.3), where F is the resonant frequency (Hz), n is the resonance number, c is a speed of sound (cm/s) and l is the length of the tube (cm) [38, 15].

$$F_n = (2n-1)\frac{c}{4l} \tag{2.3}$$

For example, the frequency of the first resonance, assuming $c = 35000$cm/s, and average male supralaryngeal vocal tract length $l = 17.5$cm. The value will be $F_1 = [2{\cdot}1{-}1][35000/(4{\times}17.5)] = 500$Hz. It is shown that as the length of the supralaryngeal vocal tract increases, the formants decrease, and viceversa. Then a speaker with a shorter supralaryngeal vocal tract will have higher resonance, and one with a longer one will have lower resonances.

In summary, we have describe that speakers anatomy influence the speech production, so the F_0 as the object of study of this thesis can characterize a speaker since the differences in vocal folds and vocal tract shape.

Chapter 3

Fundamental Frequency Estimation

Fundamental frequency is the acoustical correlate of the rate of vocal folds vibration and it is directly proportional to it [38, 15, 17, 21]. It is one of the most important *prosodic features* and known to be controlled voluntarily by the speaker [38]. The estimation of F_0 has been a mayor research topic for many decades and has resulted in numerous estimation methods. This chapter, describes the basic components for the extraction of F_0, the relation of F_0 estimation with the voicing detection problem and categorization of the different methods to extract F_0 from a speech signal.

3.1 Problem definition

It is reported in literature that F_0 or *pitch* represents a feature that can be used for identifying speakers and discriminating speaker groups (male, female and children). Reliable determination of F_0 is a difficult task. Many *pitch determination algorithm* (PDA) [17] have been developed over the years but they have been successful down certain conditions, such us, the media of the source recording and environment. Also some estimation procedures present estimation errors, such as, doubling and halving and for that reason the research to find a useful estimation of the F_0 still an open problem.

F_0 estimation is considered along with the task of *voicing determination* which refers to the classification of the frames into voiced and unvoiced ones. This is performed in long segments, on the order of 30-40 ms [2]. These two tasks tide together, thus in order to define the voiced segments from which a F_0 measurement

can be obtained, is necessary to detect the voiced segments. Glottal vibration may show aperiodicities due to changes in amplitude, rate or glottal shape waveform, or in intervals where the glottal pulses occur without an obvious regularity in time or amplitude like in glottalizations, vocal creak or fry [16]. These factors make the extraction of F_0 a complex task.

Voicing determination

The voicing determination task consist in classifying the speech segments into voiced, unvoiced, mixed excitation and silence [17]. The *voice determination algorithms* (VDA) [17] classify the voice source operation, if the vocal folds vibrate during a segment of speech then the segment is *voiced*.

F_0 estimation algorithms can be also applied to unvoiced frames for obtaining some "F_0" value. In that case, voicing determination could be implemented after the F_0 estimation as part of the postprocessing operation. Thus it can be used as a criteria to segment the signal into definitely voiced parts and in those which cannot be known for sure. When the frame contains both voiced and voiceless excitation, it is assigned a *degree of voicing* measurement, which is the total energy of the signal distributed in voiced and voiceless sources.

VDAs are used as threshold analyzers, as classifiers in pattern recognition approaches, incorporated in fundamental frequency estimation algorithms, and often implemented in conjunction with the PDA. Voicing determination and F_0 estimation are interrelated problems and both are treated separately.

3.2 Fundamental frequency estimation

The basic task is to extract the F_0 of a speech signal, known as the *first partial* or the *first harmonic* of the signal. As was mentioned, in a periodic waveform, the partials are harmonically related, meaning that they are related to the frequency of the lowest partial by integer multiplies [17, 13, 21]. The *fundamental period* (T_0) is the segment of the signal between two successive glottal pulses, which is the beginning and closing phase of the glottis. Knowing the fundamental period, the F_0 can be then computed as $F_0 = \frac{1}{T_0}$. In the analysis of a speech signal, a period can be extracted by a window function whose length should be approximately T_0.

A PDA usual division contains three main blocks: *preprocessor*, *basic extractor* and *postprocessor* [17] (see Fig 3.1).

Figure 3.1: Block diagram of F_0 estimation algorithm

The main task of the preprocessor is data reduction to increase the facility of F_0 extraction. The basic extractor performs the main task of measurement, which means to convert the signal into a sequence of local F_0 estimates. The postprocessor performs correction, error detection and smoothing of the contour.

3.2.1 Categorization of Pitch Determination Algorithms (PDA)

The categorization of PDA's can be according to their domain of operation, defined as the domain of the input signal to the the basic extractor [17]. If there is a time-domain signal that has the same time domain as the input signal, the PDA operates in *time domain*. If the input is a correlation function, Fourier spectrum, or some function derived from these, the PDA is said to operate in *spectral domain*. The common feature of spectral domain extractors is a *short-term transformation* included in the preprocessor. For this reason they are also called *short-term analysis PDA*.

Time domain PDA

In time domain PDA, the output signal of the basic extractor is a series of laryngeal pulse estimates called *pitch markers* or just *markers*.

The preprocessor consists of a filter which performs data reduction. This PDA assumes the local definition of the fundamental period (T_0) and allows the signal to be processed by period.

The basic extractor is a *threshold analyzer*, that outputs the occurrence of threshold crossings (positive, negative) of the input signal. It measures a period length that locates the exact instant of each threshold crossing and indicates the individual period boundaries by markers.

There are numerous approaches to the time domain pitch determination algorithms, some mentioned in [17]: *multichannel analysis, structural analysis, structure simplification, extraction of fundamental harmonic.*

Short-term analysis PDA

In short-term analysis PDA, the output of the basic extractor is an estimate of the local average pitch period within each frame. The preprocessor has its important step in the short-term transform by which the time-domain is left and it follows more global definition of T_0 as an average estimation per frame. The extractors for these PDAs are called *peak detectors*.

In this thesis, we concentrate in the short-term analysis algorithms, due to our interest of obtain a value for F_0 within frames, considering their independency of phase relations between the periods of the signal, and the insensitivity to phase distortions [17]. This approach is important since the focus is in the periodicity of the highest peak, that makes it resistant to noise and signal degradation.

3.3 Short-term analysis pitch determination

Hess [17] suggests a division of short-term PDAs into three main groups:

1. Correlation techniques
    - Autocorrelation
    - Distance functions (AMDF), "Anticorrelation"
2. Frequency-domain analysis
    - Direct harmonic analysis
    - Multiple spectral transform (cepstrum)
3. Time domain
    - Periodicity estimation (maximum likelihood)

The operations in these algorithms are similar (see Fig. 3.2). After an optional preprocessing operation, the signal is divided into fixed-length segments called *frames*, which have a duration of 20 - 50 ms. For frequency domain analysis, the range duration can be defined by the number of points of the signal to which the *discrete Fourier transform* is applied [15]. After segmentation, short-term transform is performed in every frame, with the objective to locate a single main peak. This peak corresponds to the F_0 estimation for the frame.

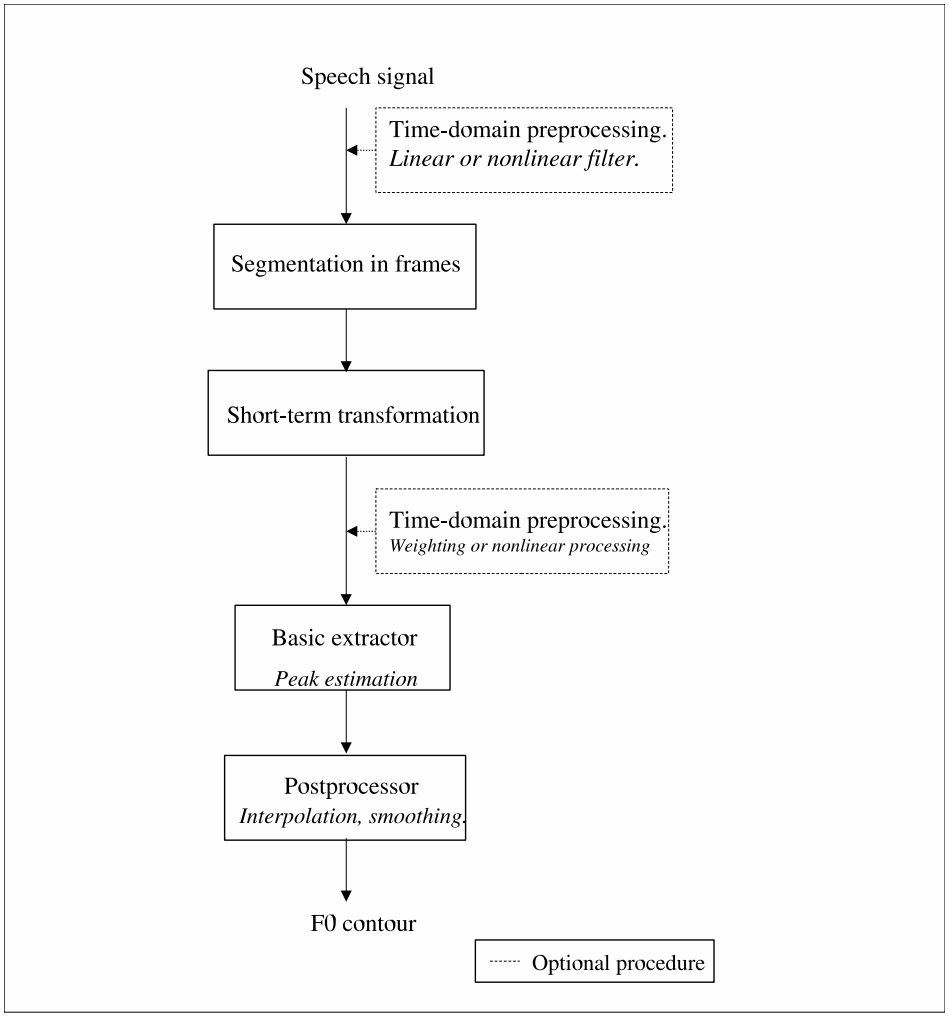


Figure 3.2: Diagram of the short-term analysis F_0 algorithm

The short-term transformation

Following the division of the signal into successive frames, they are processed individually to calculate a F_0 value. The result is an estimate of the average period length within that frame. For these algorithms, at least two periods should be located in the frame, otherwise there is no periodicity in the frame.

Most of the computing effort is on the short-term transform, which is usually a matrix multiplication of the signal vector $\mathbf{x}$ by the transform matrix $\mathbf{W}$:

$$\mathbf{X} = \mathbf{W}\mathbf{x} \tag{3.1}$$

where $\mathbf{X}$ represents the short-term spectrum of the transformation given by the matrix $\mathbf{W}$. The number of multiplications is N^2, where N is the length of the vector $\mathbf{x}$.

To reduce the number of calculations, most algorithms perform a method related to the particular transform. For instance, in some cases, the matrix **W** is decomposed in a chain of partial matrices, and the number of operations are reduced to one per row and column, as in the fast Fourier transform (FFT) [15, 17].

The idea of the transform is to identify the periodicity indicators, by focusing in one peak, maximum or minimum depending of the transform. This peak can be detected by the basic extractor. If periodicity is present, there will be a strong indication of a F_0 value, which makes this category of algorithms reliable, also in noisy environments [17].

Many algorithms are successful to approximate the F_0 value under certain conditions and type of signals [13]. Thus combining the results of various PDAs could improve the final result. Using the value of various algorithms per frame could lead to a better estimated value for the F_0, specially in the segments that some algorithm may fail and the others work.

Chapter 4

Short-term fundamental frequency estimation

The F_0 estimation algorithms process the speech signal on a frame-by-frame basis. In this chapter, we describe the following short-term analysis algorithms: *autocorrelation function* (ACF), *average magnitude difference function* (AMDF) and *cepstrum analysis*. Their main property is the short-term transform as part of the preprocessing procedure. AMDF and ACF algorithms use correlation techniques and cepstrum analysis is part of frequency domain analysis [17].

Periodic function and fundamental frequency estimation

An important property for the estimation of F_0, is the periodicity of the signal. Formally defined, a function $s(x)$ is said to be *periodic* with period p if [46]

$$s(j) = s(j + np), \tag{4.1}$$

where $s(j)$ is the jth sample for the discrete signal, for some $p > 0$ and all $n \in \mathbf{Z}$. Autocorrelation function itself is periodic. It has a global maximum for $p = 0$, if there are global maxima different than zero then the signal is periodic at lag or period p, and will have maximum values at integer multiples of p [4].

4.1 Autocorrelation function

Correlation coefficient is a measure of the similarity or the degree of linear relationship between two input functions or variables [45]. It is one of the most popular used

methods for F_0 estimation. ACF is defined as follows [13, 17, 27]:

$$r(n) = \frac{1}{N} \sum_{j=1}^{N-n-1} s(j)s(j+n), \tag{4.2}$$

where n is the *lag* or *delay* between the instantaneous and the delayed signal. The function (4.2) measures the correlation between the waveforms of the same signal at different time intervals. The autocorrelation function is periodic and shows a maximum value for those intervals of time (lags) where a period is identified. The independent variable of ACF is time and it is called *autocorrelation lag* or simply *lag* [17]. The first peak in the autocorrelation function identifies the lag that is the period of the waveform [13, 38]. Figure 4.1 shows the autocorrelation function for a voiced frame.

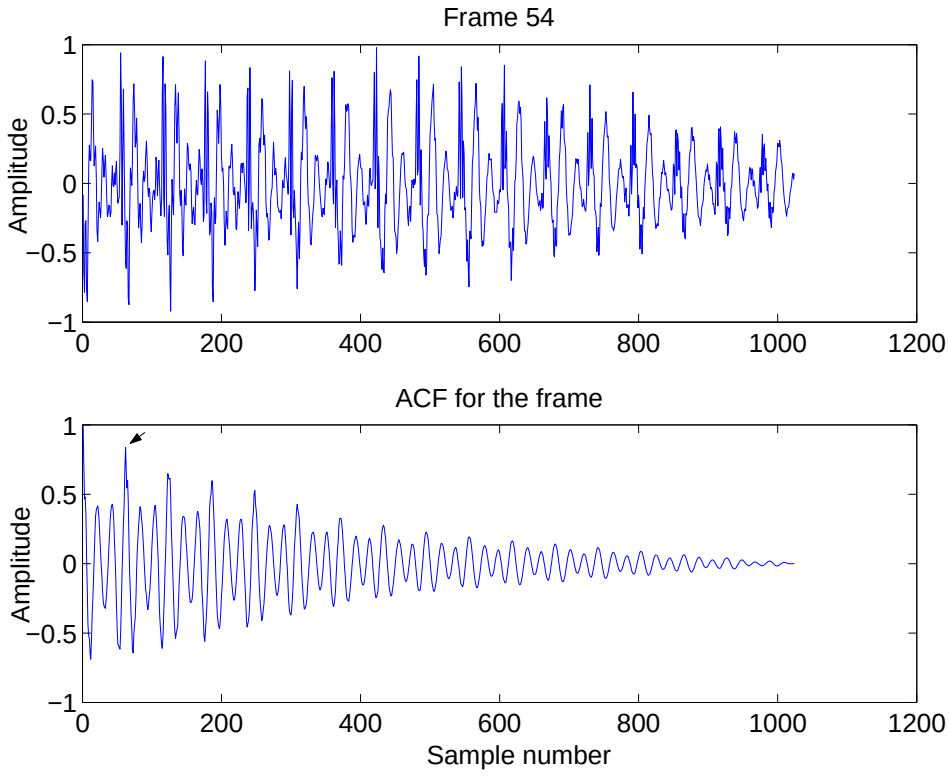


Figure 4.1: Autocorrelation for voiced frame

For quasi-periodic signals there will be a similar significant peak at $n = kT_0$ with $k = 1, 2, ...$

For ACF, the main task of the basic extractor is to identify a significant peak at $n = period$, where *period* is the T_0 duration. This period duration is expected

to be a significant peak in the range of possible F_0 values. Figure 4.2 shows the waveform of 4 seconds of a speech signal of a male speaker and the maximum values for autocorrelation in each frame.

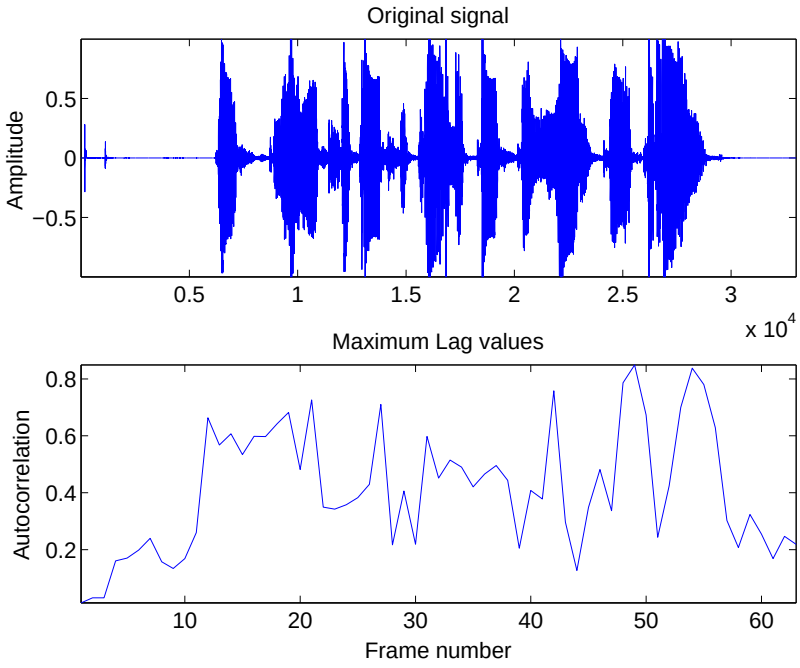


Figure 4.2: Maximum lags in autocorrelation (sampling rate 16 kHz).

We can observe that autocorrelation have maximum values distinguishable from voiced and unvoiced frames. For the frame, the maximum values for autocorrelation are at the lags related to the multiples of the fundamental period. Many F_0 estimation algorithms based on ACF classify the frames into voiced and unvoiced ones by defining a threshold for the ACF peak per frame. If the highest peak in the frame does not reach the threshold value, then is classified as unvoiced and not considered in the F_0 estimation.

Many F_0 estimation algorithms based on autocorrelation follow the scheme shown in Fig. 4.3. There are different options for the preprocessing block. According to some authors [4, 9, 17], the speech signal can be almost unprocessed, or just applying a low pass filter with a cut-off frequency of 800 Hz to reduce the influence of the higher formants.

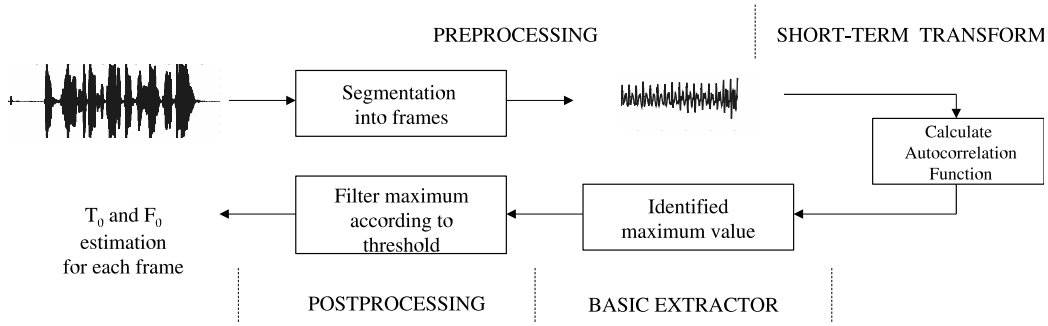


Figure 4.3: Diagram of F_0 estimation based in autocorrelation

In this work, we divide the signal into frames with the *boxcar* (rectangular) or with the *Hamming* window. We define a range to search for the fundamental period in the frame, and we define the length for the frame to be equal to two periods of the high limit in that range. Every frame is processed by the Algorithm 1, and we obtain a sequence of F_0 values for the whole signal.

Algorithm 1 Autocorrelation $(frame, fs, minlag, maxlag)$

$[r, lags] \leftarrow$ Calculate cross-correlation ($frame$) using (4.2)

$[r_{max}, t_x] \leftarrow$ maximum ($r\{minlag, maxlag\}$)

$F_0 \leftarrow fs/(minlag + t_x - 1)$

4.2 Average Magnitude Difference Function

The average magnitude difference function algorithm (AMDF) or "*anticorrelation*", is defined as follows [17, 39, 47].

$$\text{AMDF}(n) = \frac{1}{N} \sum_{j=m}^{m+N-1} |s(j) - s(j+n)|, \tag{4.3}$$

where N is the frame length, m is the starting sample of the frame, n is the lag or delay, and $s(j)$ are the samples of the speech signal. It is based in the global distance between two functions, in this case, the signal and itself shifted by n seconds. The ACF correlates the input speech at various delays, while AMDF take the magnitude difference between the delayed speech and the original signal.

The AMDF(n) is obtain by the substraction of the shifted waveform from the original one, and the sum of the magnitudes of the differences between them. The AMDF is expected to have a minimum when the lag n corresponds to T_0 (see Fig. 4.4). In the case of perfectly periodic signal, the minimum value is zero [17, 39].

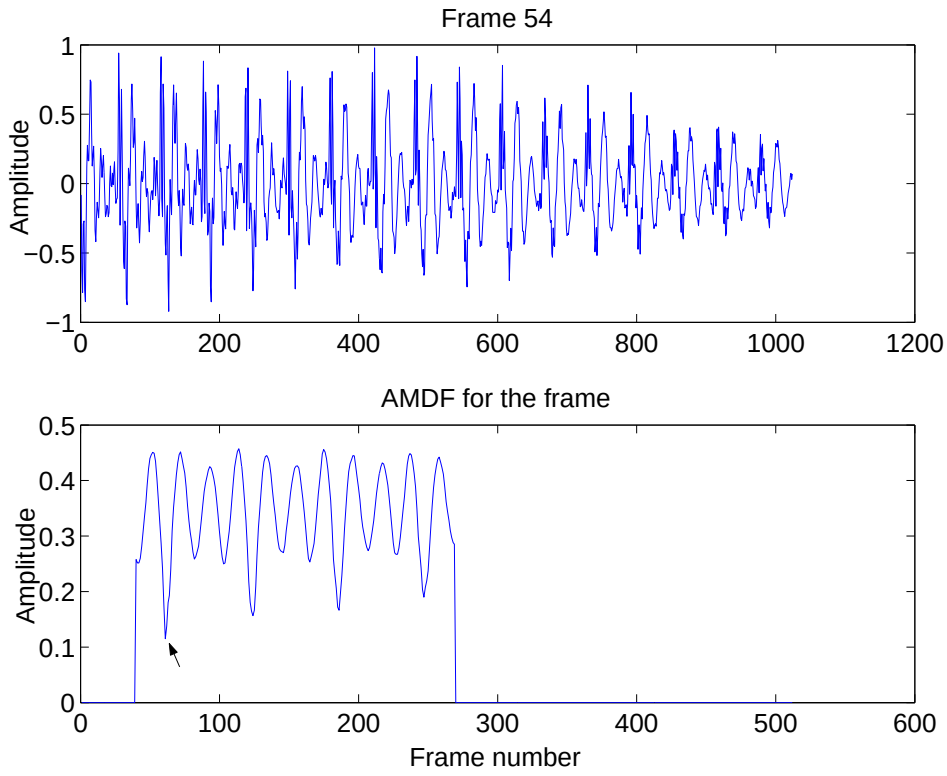


Figure 4.4: AMDF for voiced frame

The AMDF does not require matrix multiplication like ACF. Figure 4.5 shows the minimum AMDF values per frame on a signal.

Hess [17] explains that this algorithm is *phase-insensitive* (the periodic signal does not need to begin the cycle simultaneously) since the harmonics are removed as an effect of the comb filter implied in the Equation (4.3). He also mentions that AMDF is sensitive to changes that influence the magnitude of the minimum at T_0 such as intensity variations, noise and low frequency signals. Unlike other short-term analysis algorithms, AMDF does not offer a direct reference for classification into voiced and unvoiced speech segments.

Some authors include a voiced-unvoiced decision procedure in the preprocessing step, which usually is a value from the zerocrossing and energy of the segments in

the signal. Then just the voiced frames will be processed for F_0 estimation. Others prefer a smoothing procedure in the F_0 contour as a postprocessing.

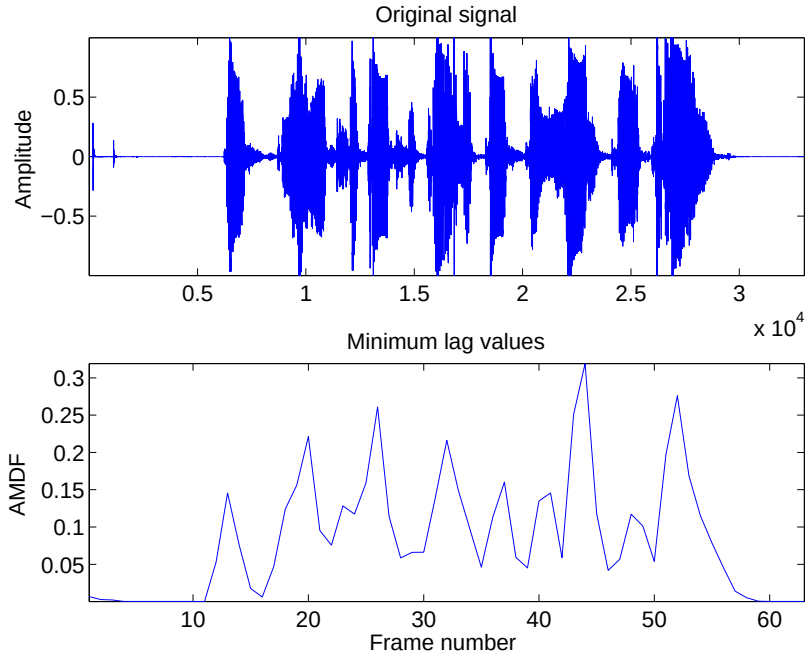


Figure 4.5: Minimum lag values for AMDF (sampling rate 16 kHz).

We include this algorithm in our study, dividing the signal into frames as in the autocorrelation algorithm. For every frame, a minimum AMDF value is identified and it is used to estimate the fundamental period, so its F_0. After the signal has been processed, we obtain a vector of F_0 values. Every frame will be processed by the Algorithm 2.

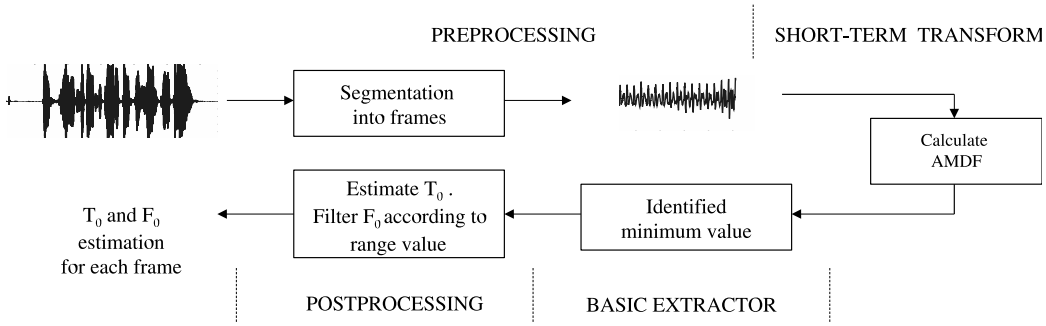


Figure 4.6: Diagram of AMDF F_0 estimation algorithm

Algorithm 2 AMDF($frame, fs, minlag, maxlag$)

```
minvalue ← infinite
for j ← minlag, ..., maxlag do
    for i ← 1, ..., length(frame) do
        dist ← |frame(i) − frame(i + j)|
    end for
    amdf(j) ← dist/i
    if amdf(j) < minvalue then
        minvalue ← amdf(j)
        T0 = j
    end if
end for
F0 = 1/T0
```

Ross *et al.* [39], give an experimental comparison between the AMDF and ACF algorithms. They stress that AMDF calculations do not require multiplications, and is more desirable for the real-time F_0 estimation. Then as the AMDF is also known as "anticorrelation", their work defines AMDF in terms of ACF with acceptable accuracy and viceversa in order to used less multiplications. It is seen in their definition that autocorrelation and AMDF are antagonist in their operation, the minimum value in AMDF is sharper than the corresponding higher peak in ACF.

4.3 Cepstrum analysis

The *cepstrum* is a common transform used for separating the excitation signal and the transfer function [17, 32]. The cepstrum is the inverse Fourier Transform of the logarithm of the spectrum of the signal. It is defined as [13, 17, 32]:

$$c(n) = \text{IDFTlog}\,|\text{DFT}s(j)|. \tag{4.4}$$

The name cepstrum comes from reversing the first letters of the word *spectrum*, which refers to the different spectral analysis done with the algorithm. The independent variable for the cepstrum is called "quefrency" which also has the first letters of *frequency* reversed. The pulse sequence in the periodic signal appears in the cepstrum as a strong peak at the quefrency lag or T_0, that is identified by the basic extractor of this algorithm (See Fig. 4.7).

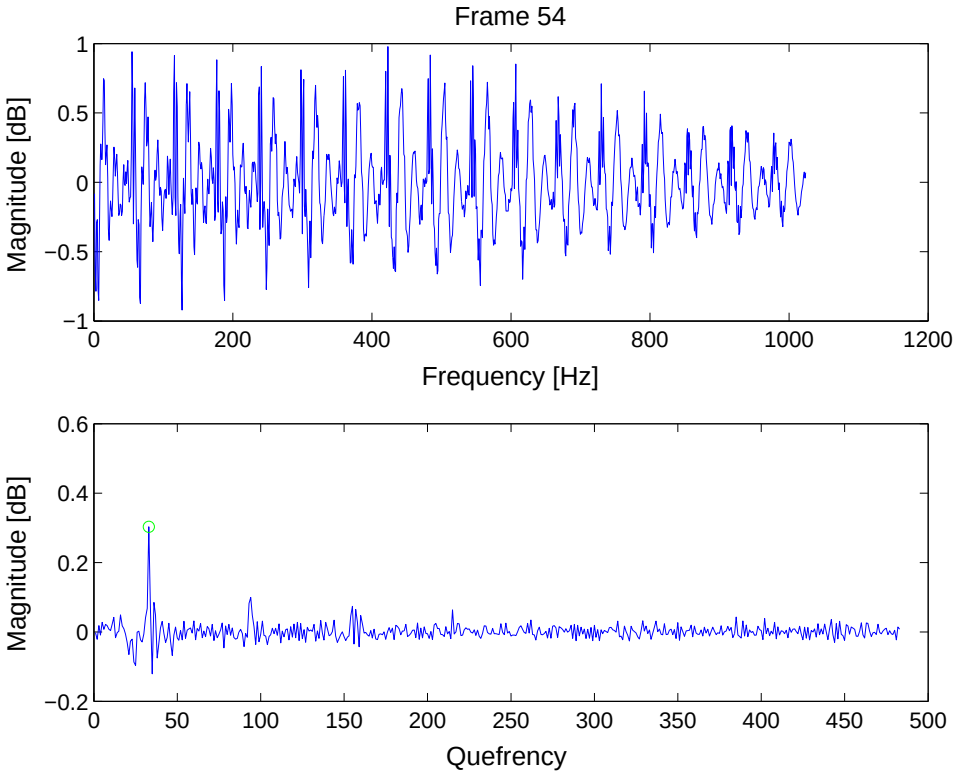


Figure 4.7: Cepstrum for voiced frame

The unprocessed signal is divided into frames that can be at 51 ms or 512 points for the FFT, and multiplied by a Hamming window. Then the discrete Fourier transform is applied to each frame. If the signal is periodic, a regular number of

peaks appear to represent the harmonic spectrum. The log magnitude is taken to reduce these peaks and translate their amplitude to an useful scale. The distance between the peaks is related to the fundamental frequency of the signal, and the highest peak indicates the quefrency related to the F_0 (Fig. 4.7).

The final step is to apply a correction procedure to adjust local errors and to identify voiced-unvoiced transitions. The most common local error is *pitch doubling* when the F_0 is estimated as the double of the true value. The procedure to define voiced and unvoiced segments with the cepstrum is to done by using a threshold for the main peak of the cepstrum in a frame. In order to overcome the estimation errors, the comparison is done in the present cepstrum and also in the cepstrum of the previous and following frames.

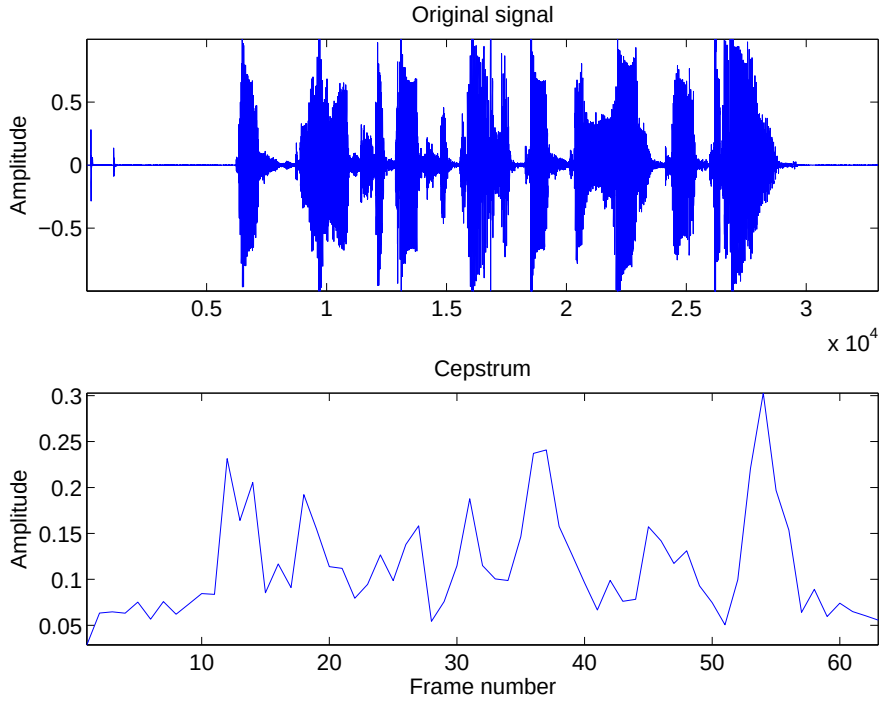


Figure 4.8: Maximum peaks for cepstrum sampling rate 16 kHz).

The cepstrum algorithm was implemented for this study (see Figure 4.9). Like in the previous algorithms, the signal is divided into frames and a Hamming window is applied. Every frame is processed with Algorithm 3.

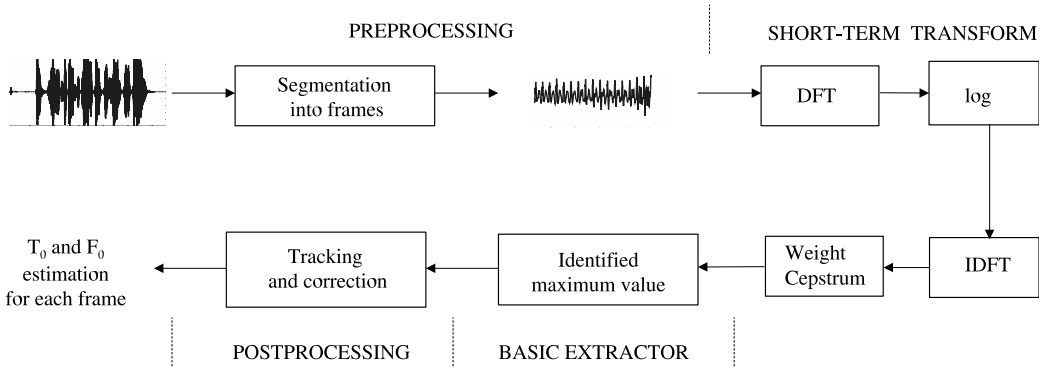


Figure 4.9: Diagram of F_0 estimation based in cepstrum analysis

Algorithm 3 Cepstrum($frame, fs, minlag$)

for $i \leftarrow 1, \ldots, length(frame)$ **do**

$logmagspectrum \leftarrow \log(abs(fft(frame(i))))$

$cepstrumcoeff \leftarrow IFFT(logmagspectrum)$

$numberofcoeff \leftarrow length(cepstrumcoeff)$

$cepstrumcoeff(0) \leftarrow cepstrumcoeff(1)$

$cepstrumcoeff \leftarrow cepstrumcoeff(2, \ldots, numberofcoeff)$

$[c_{max}, t_x] \leftarrow \max cepstrumcoeff$

$F_0 \leftarrow fs/(minlag + t_x - 1)$

end for

Chapter 5

Fundamental frequency for speaker recognition

Previously we have presented the characteristics of speech and the importance of F_0 to characterize it. We have also discussed the algorithms for estimating F_0.

Speaker recognition systems do not often include prosodic features in the identification and verification tasks, although, some techniques have incorporated F_0 and voicing information to improve performance of the system [11, 18, 6, 1]. Fundamental frequency have not been used alone as a feature for speaker recognition, due to the difficulty for current algorithms to estimate its value. However, some studies emphasize the advantages of F_0 for its robustness to noise and channel distortions compared with spectral features [11, 18].

In this chapter, we present a model for the F_0 distribution with the aim to incorporate it as an additional feature for speaker recognition. The model includes analysis of the statistics from the distribution and its representation with histograms, follow by its combination with spectral features in score-level fusion.

5.1 Speaker recognition as a pattern recognition problem

Speaker recognition has two main components: *feature extraction* and *classification* [5, 36]. In feature extraction the signal is analyzed to obtain the characteristic patterns that represents the speaker. The classifier uses the features extracted, either to identify or verify the speaker. This is done by creating a model from the features extracted during the training phase.

In the testing phase, the classifier associates a measurement to the unknown feature vector, in reference to the stored models. The model that obtains the best score gives the identity of the speaker.

One of the most commonly used feature set is *mel-frequency cepstral coefficents* (MFCC). In this study, we have include it in combination with the F_0.

Pattern matching is the step in which a match score is computed between the stored model and the distribution of the unknown speaker. The match score quantifies the similarities between the feature vectors and the models stored in the training phase [5, 36]. The modeling is performed with one or many algorithms and can be divided into *stochastic* (*parametric*) and *template* (*non-parametric*) models [21, 38]. For stochastic models, the pattern matching is a measure of the conditional probability of the observation, given the model. The template models are based on the distances measure between the observation and the model, assuming the observation to be imperfect copy of the template. The distance measure is the most intuitive method and for template model can be dependent or independent of time [5]. Reynolds [37] mentions the desirable attributes of a speaker model: (1)*theoretically supported with evidence*, (2) *generalizable to new data*, (3) *inexpensively in size and computation*. The different types of models have some or all of these attributes.

Classification or *decision making* refers to either *accepting* or *rejecting* a speaker (verification task) or identifying the target speaker (identification task).

To summarize, after obtaining the feature vectors from the test samples, the classifier calculates the match score that will be used to identify the speaker. For this, we need a model to characterize the speakers builded in the training phase and then used in the identification phase in classification. The match score comparison is performed with all the speaker models to find the one that is most similar to the unknown speaker.

In the following sections we focus on modeling and matching of the F_0 distribution.

5.2 Modeling the F_0 distribution

In this thesis, we concentrate on the F_0 distribution modeling, which means that no temporal features are considered. Temporal properties of F_0 depend more or less on the text content and the static properties reflect more the physical rather than the learned characteristics of the voice source. The assumption is that the density function of F_0 for a speaker is *almost* the same for long speech segments. For the

long-term distribution of the F_0, we use histograms as the graphical representation of the probability density. The histogram shows the number of times a particular value of the F_0 occurs in the sequence, or mostly referred as *the F_0 frequency of occurrence* [19, 38]. The shape of the histogram is important for the speaker modeling, therefore to determine the optimal number of histograms *bins* that better describe the utterance of the speaker. More bins represents more parameters to describe the distribution. Figure 5.1 shows an example of a speaker's F_0 distribution.

In this study, we consider both parametric and non-parametric models.

5.3 Parametric model

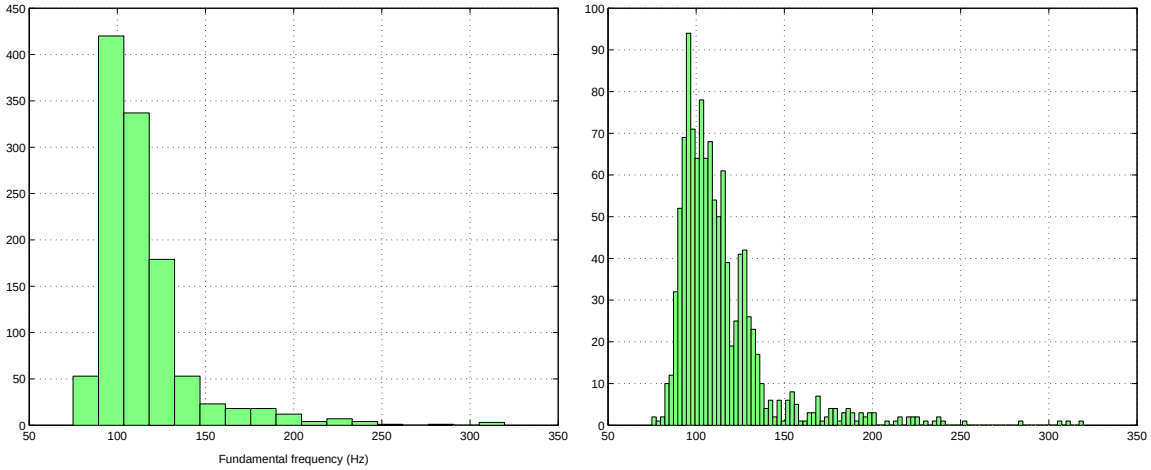


Figure 5.1: Examples of F_0 histograms for the same speaker and different sizes. Left: 17 bins, right: 100bins.

Statistical parameters associated with F_0 in long speech segments have been used for speaker recognition, for example in forensic systems [38]. We construct the *parameter* feature vector for the sequence of x_i (F_0 value for ith frame) consisting of the properties that describe and quantify the distribution: *mean*, *standard deviation*, *skew* and *kurtosis*.

Mean, as a statistical property of a distribution, is the quantity that specifies the average value, and it is computed as [38, 35]

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{5.1}$$

where N is the number of voiced frames.

Standard deviation is a quantity of how the values are spread around the mean value, also known as the second moment around the mean. Its value is the distance from each observation to the mean [38, 35]:

$$\sigma = \frac{1}{N-1}\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2} \tag{5.2}$$

Rose [38] mentions that it has been proposed that the range of the F_0, or *compass*, could be twice the standard deviation in the extremes of the mean value, because in an almost symmetrical distribution, this range will include around 96% of all the observations. This value is specially important when trying to model the distribution to a normal curve.

Skew is the quantity of the asymmetry of a distribution. It compares the amount of higher and lower frequencies shown at the extremes of the distribution. This asymmetry is *positive* if the number of higher values is more than lower values, and *negative* if the distribution contains more lower values than higher ones, or *zero* for a symmetrical distribution. The skewness of a distribution is calculated as [38, 35]:

$$\text{skewness} = \gamma_1 = \frac{\sum_{i=1}^{N}(x_i - \mu)^3}{N-1} \Big/ \sigma^3 \tag{5.3}$$

$$\gamma_1 \begin{cases} \text{Negative}, & \text{if } \gamma_1 < 0 \\ 0, & \text{symmetrical distribution} \\ \text{Positive}, & \text{if } \gamma_1 > 0 \end{cases}$$

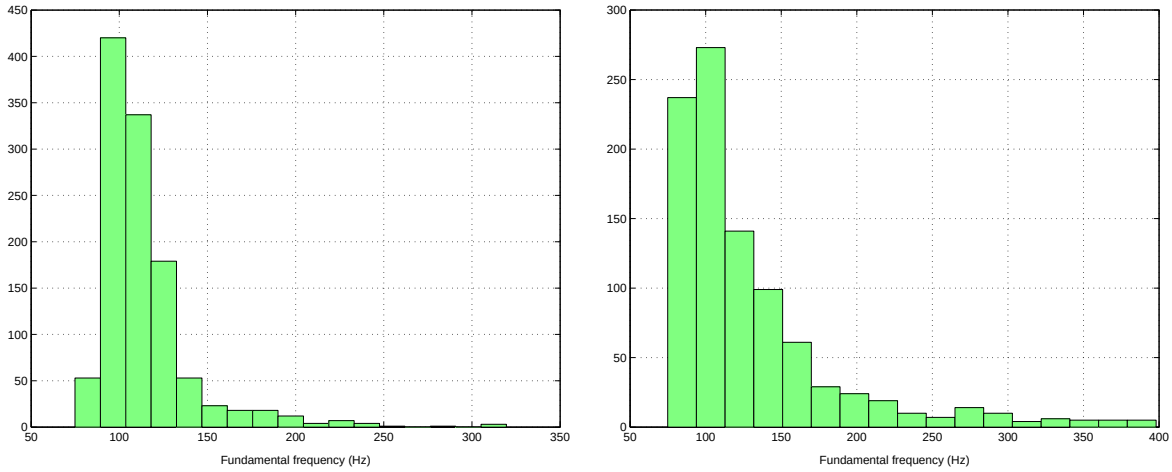
Kurtosis is the degree of "peakedness" of a distribution. It is defined as the fourth central moment of a distribution. A distribution with a high peak is called *leptokurtic*, a flat-topped curve is called *platykurtic*, and the normal distribution is called *mesokurtic* [38, 35].

$$\text{kurtosis} = \gamma_2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^4}{N-1} \Big/ \sigma^4 \tag{5.4}$$

$$\gamma_2 \begin{cases} \text{Platykurtic}, & \text{if } \gamma_2 < 3 \\ 0, & \gamma_2 = 3 \\ \text{Leptokurtic}, & \text{if } \gamma_2 > 3 \end{cases}$$

The F_0 parameter vector is then defined as $P = (\mu, \sigma, \gamma_1, \gamma_2)$. Some examples of distributions showing these statistical properties are shown in Table 5.1 and Fig. 5.2.

In the matching phase, we compare the test sample to its speaker model. This is obtained by computing a similarity measurement or match score between the parameter vector of the unknown speaker and the stored models. The score is given by a distance measure d, that is calculated with the *Euclidean distance*.



(a) Speaker 1106 (b) Speaker 1231

Figure 5.2: Histogram for test samples

Table 5.1: Example of profile feature vector

Speaker	*Sample*	*Mean*	*Standard deviation*	*skew*	*kurtosis*
1106	Test	114.06	27.361	2.85	14.97
	Train	111.38	29.114	3.051	17.847
1231	Test	129.82	57.279	2.2581	8.5618
	Train	139.03	57.674	2.7227	11.085

Euclidean distance is the difference distance between the F_0 distribution of the unknown speaker and the distribution of the model. The Euclidean distance for two distribution $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{5.5}$$

For the parametric approach, the smallest difference relates the speaker to its model.

5.4 Non-parametric model

With the sequence of features vectors obtain from the speech samples, we proceed to classified to which speaker the utterance belongs.

To compare the test sample to its speaker model, for the non parametric model, we compute a similarity measurement. We study both, the Euclidean and *Kullback-Leibler*(KL) distances. The Euclidean distance is calculated as defined in the previous section.

Kullback-Leibler distance (relative entropy) is the measure of the distance between two distributions [5, 7, 19], in this case between the F_0 distributions of the model and the unknown speaker. The Kullback-Leibler distance of two distributions p_k and q_k is defined as:

$$\mathrm{d(p,q)} = \sum_k p_k \log \frac{p_k}{q_k}, \tag{5.6}$$

where p_k and q_k are the density functions for their respective distributions. In general, $d(p, q) \neq d(q, p)$. The Euclidean distance represents a metric between the vectors, while Kullback-Leibler is a distance between the probability functions.

For the shape of the distribution, the result of the classification phase are the measurements obtained by the Euclidean and Kullback-Leibler distances between the F_0 histograms for the test and the model of the training samples for all speakers. A smaller distance suggest the speaker's sample is closer to the model speaker.

At this point, where both parametric and non-parametric model are classified, a decision about the identity of the speaker can be made based on F_0.

Figure 5.3 shows the F_0 matching phase for two distributions and their respective similarity measurements. It is noticed that the matching of test and train samples from the same speaker have smaller distances than when the distributions correspond to different speakers.

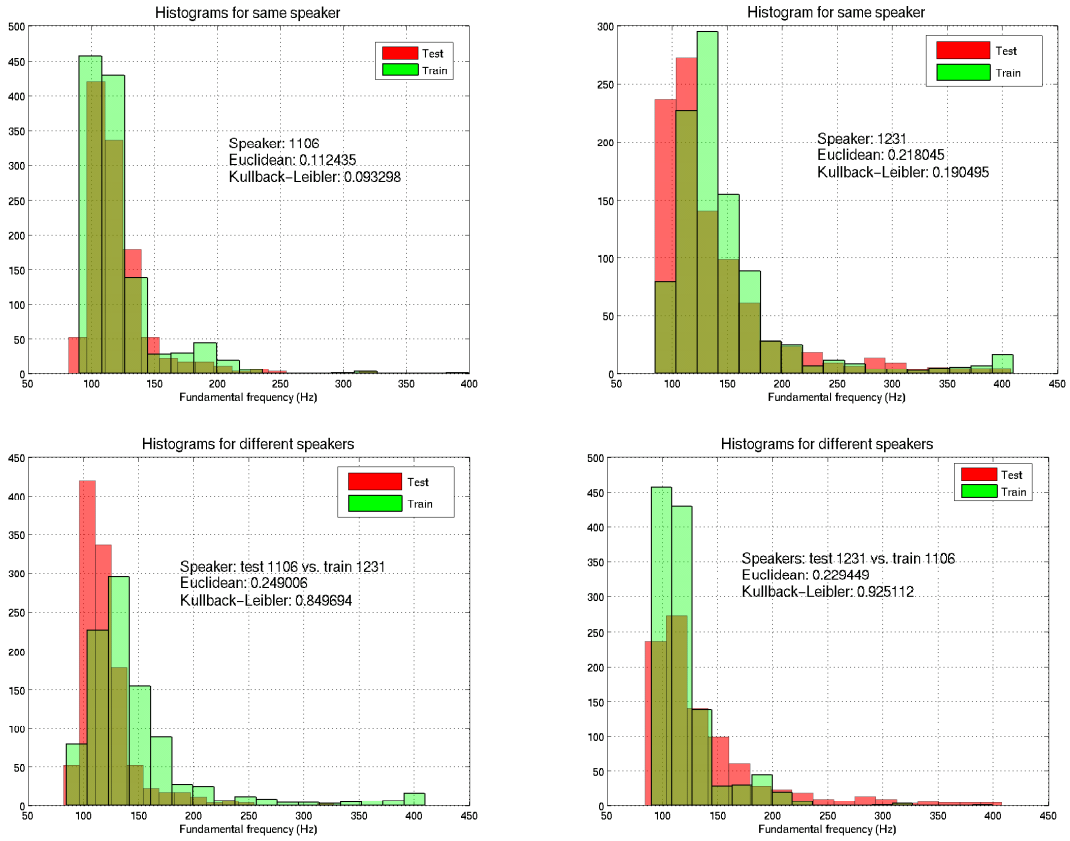


Figure 5.3: Examples of F_0 histogram matching (histogram size = 27 bins).

5.5 Mel-Frequency Cepstral Coefficients

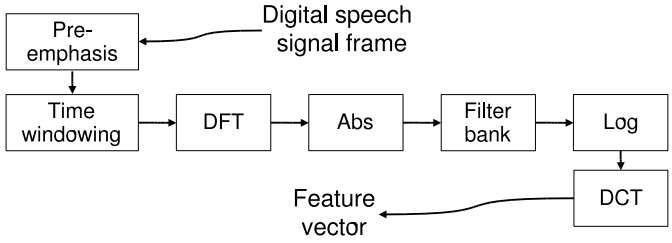


Figure 5.4: Scheme of MFCC signal processing steps [41]

MFCC is the most commonly used set of features in speaker recognition. It is the parametric representation of the speech signal based on the Fourier spectrum. The process to compute it is:

1. Division of signal into overlapping frames.

2. Pre-emphasis of the signal, which consists on rising the level of higher frequencies [15].

3. Every frame is multiplied by a window function, such as Hamming window.

4. Determination of the amplitude spectrum using FFT.

5. Conversion to Mel scale by applying a filter bank to the magnitude spectrum.

6. Application of the discrete cosine transform (DCT) to the logarithm of the filter bank output.

The first coefficient is a measure of the energy in the signal and depends on the intensity [11, 23]. The rest of the coefficients contains the information of the vocal tract filter and are fairly uncorrelated [21]. MFCC is known to deconvolve the source and the vocal tract. However, it is noticed that in practice the cepstrum coefficients are affected by high pitched voices [11]. For more details about mel-scaled cepstral coefficients, see [8, 23, 41].

5.6 Classifier fusion

The *classifier fusion* is used for combining two or more models, with the aim to improve accuracy over the best individual classifier. The results of every classifier gives an "*expert*" opinion of the identity of the speaker based in the match score between the unknown speaker and the model. There are many fusion techniques for two or more classifiers: *sensor level*, *feature-level*, *decision-level* and *score-level* [23]. In this study, we combine the features in score-level (See Fig. 5.5) [22].

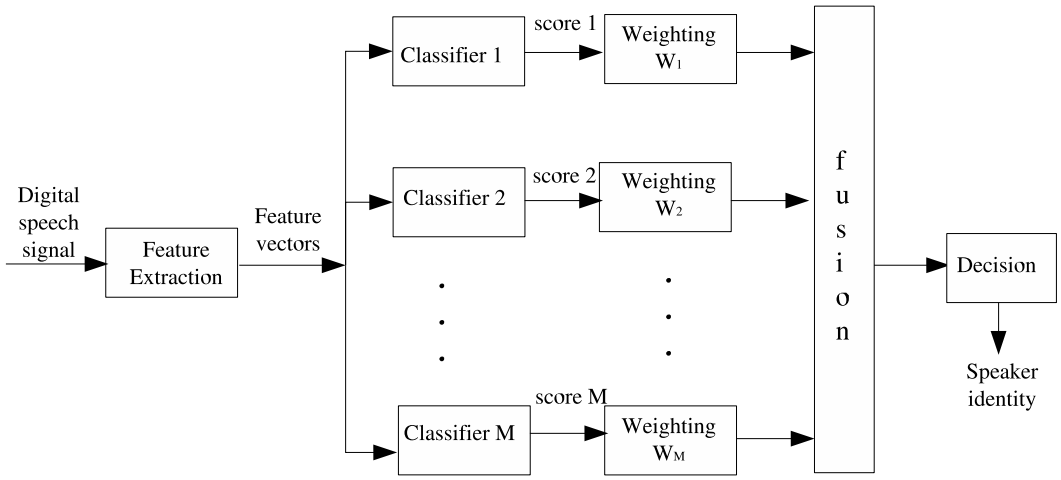


Figure 5.5: Scheme of score-level fusion for one feature.

In this approach, every feature set is modeled separately and has a specialize classifier to output the scores that will be combined to take a decision. The match scores are combined with a *fusion rule*. We use *weighted summation* as the fusion rule, defined as follows:

$$\text{Score}(j) = \sum_{i=1}^{M} w(i)\text{Score}(i,j) \tag{5.7}$$

where M is the number of classifiers in the fusion, Score(i,j) is the score of the classifier i for the speaker j, and $w(i)$ is the weight for the classifier i. An important point to consider is the normalization of the scores and weights in the range $[0, 1]$ so that $\sum_{i=1}^{M} w(i) = 1$ and $\sum_{i=1}^{M} \text{Score}(i,j) = 1$, for all i. The classifier with more weight is considered more important for the recognition process. In our case, the identification decision is given by the smallest distance calculated after fusing all the classifiers, as:

$$D_{\text{combined}} = \sum_{i=1}^{M} w(i)\text{Score}(i,j) \tag{5.8}$$

$$\text{Decision} = \arg\min(D_{\text{combined}}), \text{for all M classifiers} \tag{5.9}$$

Chapter 6

Experiments

We present the experimental setup and results incorporating F_0 for speaker recognition. The F_0 was estimated with the PRAAT application [34] which computes F_0 using the autocorrelation algorithm described in [4]. For the data set considered, approximately the half of the number of frames were classified as voiced and that gives enough information about the F_0 long-term distribution.

The F_0 tracking is a problem studied in great extend over the past years. We implemented the algorithms described in Chapter 5 which revealed a lot of challenges in estimating both the F_0 and voicing degree. In this thesis, we consider F_0 modeling in speaker recognition, and assume that the estimation is already done accurate as possible. For this reason, we decided to take PRAAT and use the existing implementation in an "off-the-shelf" manner.

For the fusion of classifiers, we used the MFCC features extracted using previously implemented tools at our department [42]. We consider *closed set text-independent* speaker identification task. We use score-level fusion by weight summation. In the following sections we describe the experimental setup, experiment procedure and the results.

6.1 Experimental setup

We use the male subset of the *NIST 1999* speaker recognition corpus [29], consisting of 230 speakers. The speech material is conversational, with mainly college students that didn't know each other. The speakers could make or receive one call per day. The received call was on the same phone line, while the originated call was required to be from different phone lines [28]. For training, we use the files with an "a"

Table 6.1: Summary of the NIST-1999 subset.

Language	English
Speakers	230 male speakers
Speech type	Conversational
Quality	Telephone
Sampling rate	8.0 kHz
Quantization	16-bit μ-law
Training speech (avg.)	59.0 sec.
Evaluation speech (avg.)	59.0 sec.

Table 6.2: Parameters for MFCC feature.

Frame length	30 ms
Frame shift	20 ms
Window function	Hamming
Model	Vector Quantization
Codebook size	64
Pre-emphasis	0.97
Number of coefficients	12
Mel filters	27

following the speaker number and the files with "b" for testing. The text content of the utterance is not fixed and varies from speaker to speaker. A conversation topic was suggested, but speakers were free to talk about different topic. The recording session for "a" and "b" files is different. Table 6.1 summarizes the characteristics of the speech material. We perform closed set text-independent speaker identification, in which the speaker is known to belong to the group of N speakers. The speakers are divided into two groups, 50 speakers (first file names in ascendant order) are used for parameter tuning for optimizing the histogram bin size and the weights of the classifiers fusion. The rest 180 speakers are analyzed with the parameters obtain in the tuning set.

For MFCC feature, the parameters were set as described in table 6.2, based on the work of Kinnunen *et al.* [22, 23, 24]. The F_0 is estimated with the parameters presented in Table 6.3 from every frame. PRAAT calculates both, the F_0 values and their time stamps, and we removed the time stamps to get the fundamental frequency values only.

Table 6.3: Parameters for F_0 feature using PRAAT.

Pitch floor	75 Hz
Pitch ceiling	400 Hz
Max. number of candidates	15
Very accurate	yes
Silence threshold	0.03
Voicing threshold	0.45
Octave cost	0.45
Octave-jump cost	0.45
Voiced/unvoiced cost	0.45

Some parameters used by PRAAT to calculate the F_0 values are of main importance for testing. Since the data set includes only male speakers, we set the pitch floor and ceiling in the range from 75 Hz to 400 Hz. Candidate values for F_0 out of this range will be ignored by the analysis algorithm. In the case of female speakers, the range could be set to 100 - 600 Hz.

The *very accurate* parameter is used to process the frame with a Gaussian window with a physical length of $6/(pitchfloor)$, which is twice the effective length for the Boersma's algorithm [4]. The rest of the parameters are used to determine the cheapest path through the F_0 candidates.

6.2 Speaker Modeling

In the training phase (See Fig. 6.1), we enroll the speaker creating a model based on the extracted features and store it in the database. In the identification phase, the matching algorithm compares the scores of the testing samples and the stored models. These results are used to make a decision on the speaker identity.

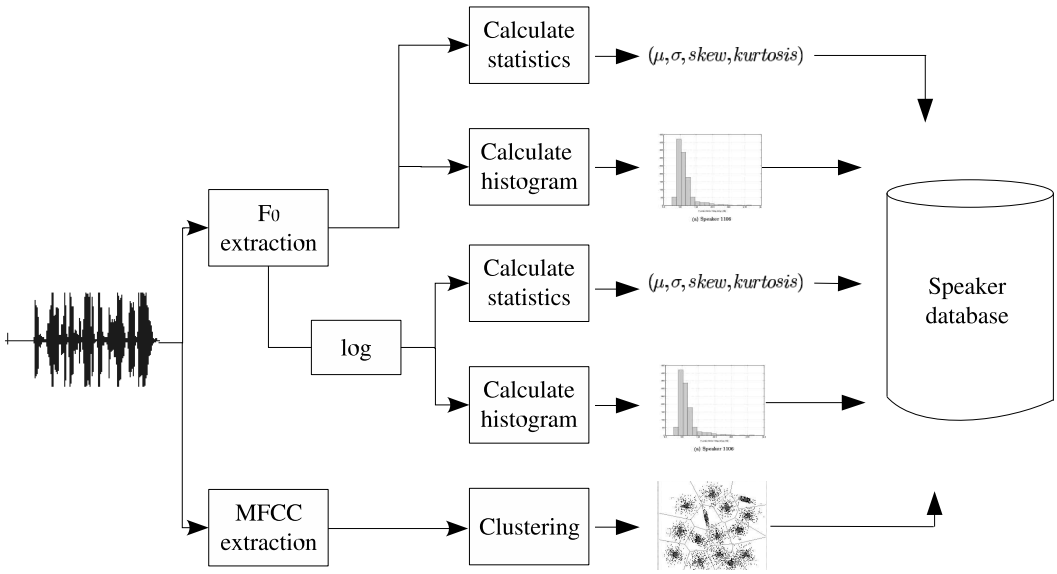


Figure 6.1: Diagram of training phase

6.2.1 Non-Parametric F_0 model

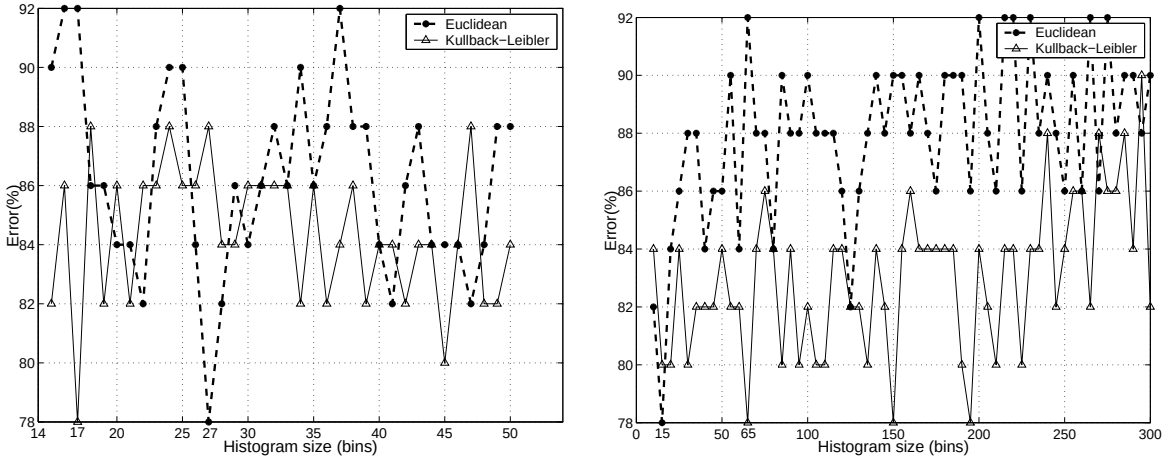
The F_0 probability density can be estimated by the histogram method. During the training phase (See Fig. 6.1), we obtain the histogram for the F_0 distribution of each speaker, and store them for the matching phase. F_0 is processed in both *linear* and *logarithmic* scale. In the linear approach, we use the estimation of the raw F_0, and for the the second approach, the logarithmic form of F_0. It has been observed in [43], that the clean F_0 has a lognormal distribution, so the logarithm of the F_0 has a Gaussian distribution.

For linear and logarithmic approaches, we perform the test calculating the histogram and the statistical properties of the F_0 for the test speaker. Then we compute the Euclidean and Kullback-Leibler distances to compare the stored model and the unknown speaker. Table 6.4 shows the number of bins that yielded to the lowest error rate during the parameter tuning phase. According to our experiments, there

Table 6.4: Histogram size (bins)

	Euclidean	*Kullback-Leibler*
F_0	27	17
$logF_0$	15	65

is no clear trend how histogram size and error rate are related. Fig. 6.2 shows the performance according to histogram size for each distance method.



(a) raw F_0 (b) logF_0

Figure 6.2: Histogram size selection

Euclidean and Kullback-Leibler distances are normalized and weighted with the objective to be used in the fusion phase, in combination with the other models (parametric and MFCC).

6.2.2 Parametric F_0 model

The statistical properties for the F_0 are effective to describe the distribution and have been used as prosodic features for speaker recognition [38, 43]. We calculate the statistical properties of the distribution in a four dimensional vector $(\mu, \sigma, \gamma_1, \gamma_2)$, and refer to it as the parametric model.

For testing, we estimate this vector and compare it with the stored model using Euclidean distance. The computed match scores are normalized and weighted as a preparation for the fusion with the rest of classifiers.

6.2.3 MFCC model

For MFCC, the model is created by clustering the feature vectors in the training phase using *K-means algorithm* [26]. The clustering result is a set of vectors called *codebook*. In our study, we use a codebook of size 64. The matching is performed by *vector quantizing* the unknown sample with the codebook [24].

Table 6.5: Error rates (%)

	Euclidean	*Kullback-Leibler*	*Parametric*	*Fusion*	*Fusion*
F_0	95	93.9	92.8	91.1	88.9
$logF_0$	93.9	89.4	93.9	91.1	

Table 6.6: Error rates (%) for noisy conditions. SNR = 10 dB

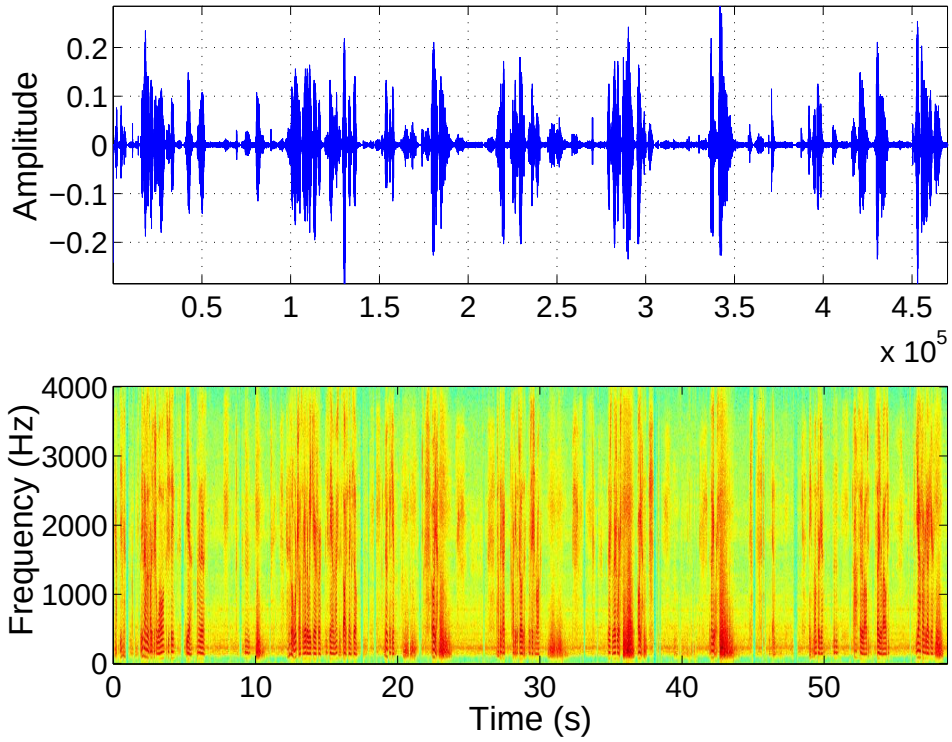
	Euclidean	*Kullback-Leibler*	*Parametric*	*Fusion*	*Fusion*
F_0	96.1	93.3	93.3	89.4	90.6
$logF_0$	96.7	89.4	95	91.1	

6.3 Results

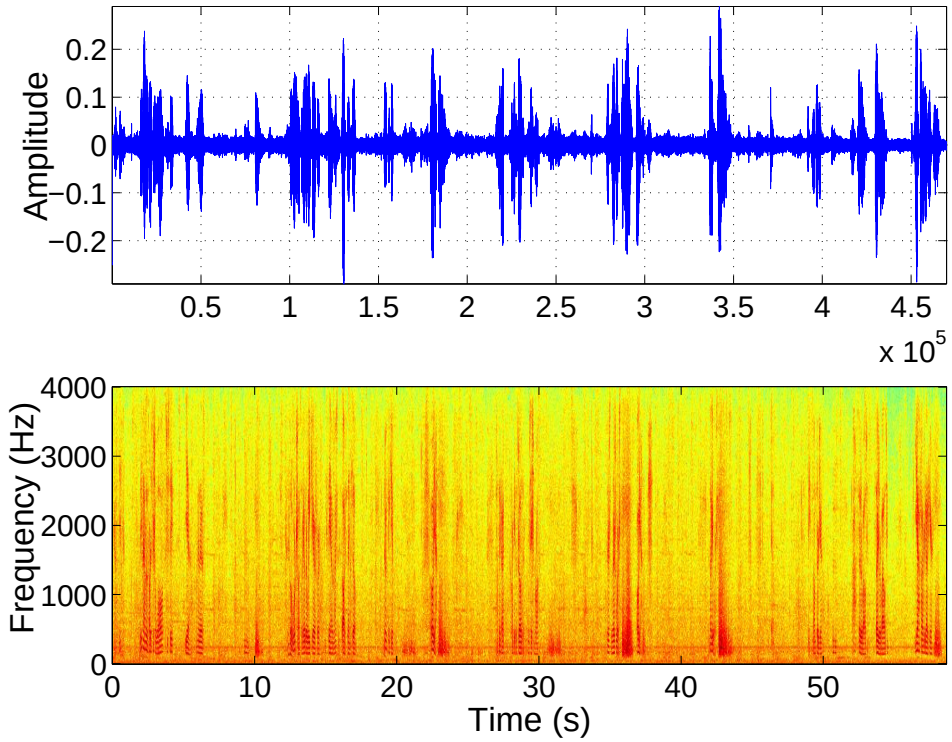
6.3.1 Results for F_0 classifiers

We tested the F_0 models and combined them to create the model that will represent the F_0 contribution to the recognition process. First, we selected the weight values for the parametric and non-parametric model based on the lowest error rate on the tuning set (See Table 6.7). Table 6.5 presents the results for the clean training and testing conditions. The error rate presented is $(N_{incorrect}/N_{total}) \cdot 100\%$, where $N_{incorrect}$ is the number of test segments incorrectly classified and N_{total} is the total number of test segments. As expected, the F_0 alone gives low recognition rate. In the linear approach, the lowest value is 92.8% obtained with the parametric model (F_0 statistics). In the logarithmic approach, the method that uses Kullback-Leibler distance was better in the matching of histograms, slightly better than the Euclidean distance (89.4% vs. 93.9%). In the fusion, the information collected from the different methods leads only to marginal improvement. Table 6.7 shows the weights selection for the combination of the six models for F_0. For linear processing, the models are considered equally useful, while in logarithmic scale, the Kullback-Leibler approach is more important.

Next, we tested the same models in noisy conditions (See Table 6.6). The original speech samples are recorded over the telephone, so is considered "noiseless" although occasional background sounds can be perceived in some of the samples. We applied additive *factory* noise [30] with a *signal-to-noise ratio* (SNR) of 10 dB. Fig. 6.3 shows a speech sample before and after adding noise.



(a) Original sample



(a) Sample corrupted by additive Factory noise (SNR = 10dB)

Figure 6.3: Waveforms and spectrograms for speaker 4402b.wav

Table 6.7: Weights for the fusion of F_0 classifiers

	F_0			$logF_0$		
Classifiers	*Euclidean*	*KL*	*Parametric*	*Euclidean*	*KL*	*Parametric*
3	0.33	0.33	0.33			
3				0.27	0.51	0.21
6	0.063	0.063	0.063	0.188	0.562	0.063

6.3.2 Combining F_0 and MFCC

Before fusing F_0 models with the MFCC model, we evaluated the correlation coefficients for the distance matrices of each classifier. The purpose of this is to investigate which classifiers are less correlated and potential for providing additional information in the fusion. The results are shown in the Table 6.8, and we can see that MFCC shows clearly the lowest correlation compared to the F_0 models. This suggests that the procedure to fused these two features is appropriate. The different measurements obtained from F_0 are more correlated between the distance approach as compared with the parametric and nonparametric approach.

The combined distance of F_0 and MFCC is defined as:

$$D = \alpha \cdot d_{MFCC} + (1 - \alpha) \cdot d_{F_0}, \tag{6.1}$$

where d_{MFCC} and d_{F_0} are the distances computed by their respective classifiers and $0 < \alpha \leq 1$ is the weight or degree of contribution from each feature to the final distance. For the fusion, we find the α that yields the lowest error rate for the combination.

Table 6.8: Distance matrices correlations

		F_0			$\log F_0$			*MFCC*
		Euclidean	*KL*	*Parametric*	*Euclidean*	*KL*	*Parametric*	
F_0	*Euclidean*	1.000	0.666	0.355	0.981	0.765	0.368	-0.030
	KL		1.000	0.259	0.683	0.844	0.235	-0.016
	Parametric			1.000	0.332	0.285	0.657	-0.124
$logF_0$	*Euclidean*				1.000	0.779	0.348	-0.034
	KL					1.000	0.267	-0.013
	Parametric						1.000	-0.118

We performed the recognition in both matched and mismatched SNR conditions. The *matched* case means that training and identification phase are done both on clean speech or with additive noise one. The *mismatched* case is when training is done on clean speech and identification on speech with additive noise, and vice versa. Table 6.9 summarizes the results for both cases.

Table 6.9: Error Rates % for noisy conditions. SNR = 10 dB

	Training	*Identification*	F_0	*MFCC*	*Fusion*	$(1-\alpha)$
Matched	clean	clean	88.9	23.9	22.8	0.02
	noise	noise	90.6	34.4	28.3	0.036
Mismatched	noise	clean	90.0	34.4	67.8	0.16
	clean	noise	88.9	91.7	86.1	0.79

For MFCC, we obtained a baseline of 23.9 % and for noisy conditions the error rate is 34.4 %.

The accuracy of the MFCC feature radically decreases under noisy conditions, whereas F_0 is relatively robust. F_0 contributes to improve accuracy in all cases. Ramachandran *et al.* [36] mention that to achieve robustness at the feature level, requires to configure them to show small variation for different conditions, since there is a great need of robust speaker recognition systems also under mismatched conditions. Then F_0 feature could contribute in these cases.

6.3.3 Samples correctly classified by F_0 but not by MFCC

In this work, we observed that some speakers were recognized by the F_0 classifier but not by the MFCC. The characteristics of these speech samples vary from speaker to speaker. For instance, some of the original recordings had background noise, such as television programs, children playing, beep of call entering, and so on. In some cases, the speaker's rhythm was different in both recording sessions, like in one laughing and the other one speaking normally. Table 6.10 shows the file name for the speaker train and test samples that were identify with F_0 feature only. An interesting observation can be seen in Fig. 6.4 where the distributions for training and identification are matched for two of the mentioned speakers. For comparative purposes the figure presents the long-term average spectra (LTAS) along with log F_0 distributions. The figure shows mismatches for the LTAS in both intensities and spectral shapes, but the log F_0 distributions are very close.

Table 6.10: Speakers identify correct with F_0 and incorrect by MFCC

Matched		*Mismatched*	
Clean - Clean	*Noise - Noise*	*Clean - Noise*	*Noise - Clean*
4202	4270	4237	4270
4402	4402	4241	4391
4633	4487	4270	4402
4785	4543	4391	4996
4949		4402	4999
4996		4487	
		4531	
		4535	
		4610	
		4621	
		4841	
		4914	
		4949	
		4999	
6	4	14	5

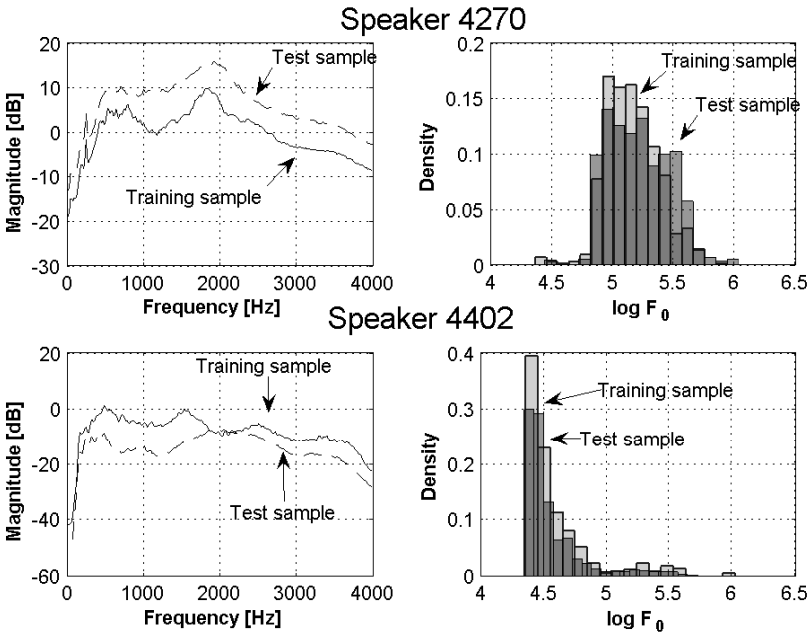


Figure 6.4: Long term average spectra (left) along with F_0 histograms (right). For two speakers recognized by F_0 but missed by MFCC

6.4 Summary of results

From the results the following observation were made:

- The best individual F_0 classifier is the log F_0 matching with Kullback-Leibler distance, in the four testing conditions it got the minimum error rate. A reason for this is that the number of free parameters (histogram size) is larger than in the other models, so the distributions can be better discriminated. The next classifier was the parametric model with the linear approach (raw F_0).

- Under noisy and mismatched conditions, the performance of F_0, although poor, only slightly varies, while MFCC decreased its performance compared to clean speech conditions which is known to perform very good. Under noisy condition F_0 is much more robust than spectral features.

- Under noisy conditions the relevance of the F_0 model (fusion of 6 classifiers) in the fusion with MFCC increases (the last column of Table 6.9).

- Some samples of speakers were recognized correctly by F_0 but not by MFCC. In these cases, the original samples are characterized by containing background noise, quality of the voice and rhythms of conversation varies between test and train sample.

- The low correlation with MFCC scores, suggest that a better combination would be possible. The fusions were done only with the weighted sum as a fusion rule.

Chapter 7

Conclusions

The fundamental frequency is considered to be an important cue to discriminate between speakers. In this thesis, we presented a study of the F_0, its estimation and use as an additional feature in speaker identification. Closed-set speaker identification was experimented in this study.

7.1 F_0 extraction

To model a prosodic feature like F_0, we need to define an appropriate extraction technique. For F_0 tracking, a very large number of algorithms have been proposed. Many of them have been designed for a particular problem, such as music, singing and voice, or have specific limitations that can be use for particular signals, recording conditions, and so on. The accurate determination of speech F_0 is still an open problem. The difficulties for the estimation lies in the extraction, where the F_0 can be estimated only for voiced speech segments. This means that the classification of voiced and unvoiced sounds is a key part of the F_0 estimation, although is not required for the preprocessing phase. It is usually implemented in the postprocessing as part of the error correction and smoothing of the contour. Also estimation errors, halving and doubling affect the extraction.

In this work, we implemented the algorithms described in Chapter 5, but after experimentation, we found out that achieving high accuracy is difficult for telephone quality samples. However, the information they provide, as F_0 and voicing degree can be used to define a speaker. An interesting future direction to the estimation problem could be to fuse the complementary results provided by different extractors, when one fails and the others work. This requires more extensive work to define the

best way to integrate their estimation, therefore to obtain a more reliable data for F_0.

7.2 F_0 modeling for speaker recognition

The incorporation of F_0 to text-independent speaker recognition was studied, with specific focus on the reliability to differentiate between speakers and its robustness under noisy conditions. We estimated training and testing distributions of F_0 and modeled them using histograms and statistic parameters. Euclidean and Kullback-Leibler distances were used for measuring the dissimilarities between the histograms (nonparametric approach). We also computed the statistical properties of the distributions (parametric approach) and compared the parameter vectors with the weighted Euclidean distance. Although, parametric and nonparametric models are calculated from the same F_0, the study shows their potential information for the model of the feature.

The experimental results indicates that F_0 feature used alone for speaker identification has limitations. The identification error rate range is 89.4 - 96.7 % for matched cases (training and testing: clean - clean, or noise - noise) and mismatched ones (noise - clean or clean - noise). Noise was added to the speech material to test the system under those conditions.

The best individual F_0 classifier is the log F_0 domain with Kullback-Leibler dissimilarity, with error rate of 89.4% in all test conditions.

MFCC feature is known to have a high performance in speaker recognition, although its limited to almost clean quality speech. In our experiments, the error rate for MFCC was between 23.9 - 91.7 %. The fusion of F_0 with the MFCC improves the identification error rate to the range of 22.8 - 86.1 %. The situation where the training material was noisy and the testing was clean, shows the relevance that F_0 can have in the identification task. It indicates that F_0 can add useful information for the recognition. This results suggest that the performance of speaker recognition system which uses spectral feature could be increased by incorporating F_0, and that the fusion of uncorrelated feature sets it is clearly better than the classifier alone. This study included the model of F_0 using only score level fusion with weighted summation. Other alternatives to model and to fused the F_0 classifiers could be tried, even including the dynamic features of F_0 could offer the opportunity to explore the improvements of adding this feature to speaker recognition.

Bibliography

[1] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey. Modeling prosodic dynamics for speaker recognition. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume IV, pages 788 – 791, Hong Kong, April 2003.

[2] B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recogntion. *IEEE transactions on Acoustics, Speech and Signal Processing*, ASSP-24(3):201–212, June 1976.

[3] K. Bartkova, D. L. Gac, D. Charlet, and D. Jouvet. Prosodic parameter for speaker identification. In *Proceeding of International Conference on spoken Language Processing (ICSLP)*, pages 1197 – 1200, Denver, Colorado, USA, 2002.

[4] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences(IFA)*, volume 17, pages 97–110, University of Amsterdam, 1993.

[5] J. P. Campbell. Speaker recognition: a tutorial. In *Proceedings of the IEEE*, volume 85, pages 1437– 1462, September 1997.

[6] Y. Cheng and H. C. Leung. Speaker verification using fundamental frequency. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, November-December 1998.

[7] T. M. Cover and J. A. Thomas. *Elements of Information theory.* John Wiley & Sons, Inc., 1991.

[8] S. B. Davis and P. Mermelstein. Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustic, speech and signal processing*, ASSP-28(4), 1980.

[9] A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *Acoustical Society of America*, April 2002.

[10] Oxford english dictionary. www page, February 2005. http://dictionary.oed.com/.

[11] H. Ezzaidi, J. Rouat, , and D. OŚhaughnessy. Towards combining pitch and mfcc for speaker identification systems. In *Proceedings of European conference on speech communitation and technology (EUROSPEECH)*, pages 2825–2828, September 2001.

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

[13] D. Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical report, Department of Computer Science, University of Regina, Canada, November 2003.

[14] The larynx and voice: Basic anatomy and physiology, February 2005. http://www.hopkinsmedicine.org/voice/anatomy.html.

[15] J. Harrington and S. Cassidy. *Techniques in Speech Acoustics*. Kluwer Academic Publishers, Dordrecht,The Netherlands, 1999.

[16] P. Hedelin and D. Huber. Pitch period determination of aperiodic speech signal. In *Proceedings ICASSP*, pages 361–364, 1990.

[17] W. Hess. *Pitch Determination of Speech Signals: algorithms and devices*. Springer-Verlag, Berlin, 1983.

[18] K. Iwano, T. Asami, and S. Furui. Noise-robust speaker verification using f_0 features. In *Proceedings International Conference on Acoustics, Speech, and Signal Processing (INTERSPEECH - ICSLP)*, pages 1417–1420, Jeju Island, Korea, October 2004.

[19] F. Jauquet, P. Verlinde, and C. Vloeberghs. Histogram classifiers using vocal tract and pitch information for text-independent speaker identification. In *ProRISC 9th Annual Workshop on circuits, systems and signal processing*, pages 213–218, 1997.

[20] M. Jiang. Fundamental frequency vector for a speaker identification system. *Forensic linguistics*, 23(1):95 – 106, 1996.

[21] T. Kinnunen. Spectral features for automatic text-independent speaker recognition. Licentiate's Thesis, December 2003.

[22] T. Kinnunen, V. Hautamäki, and P. Fränti. On the fusion of dissimilarity-based classifiers for speaker recognition. In *Proceedings 8th European conference on speech communitation and technology (EUROSPEECH)*, pages 2641–2644, Geneva, Switzerland, September 2003.

[23] T. Kinnunen, V. Hautamäki, and P. Fränti. Fusion of spectral feature sets for accurate speaker identification. In *Proceedings 9th International Conference Speech and Computer (SPECOM)*, pages 361–365, St. Petersburg, Russia, September 2004.

[24] T. Kinnunen, E. Karpov, and P. Fränti. Real-time speaker identification and verification. *IEEE transactions on speech and audio processing*, 2005. Accepted.

[25] Anatomy and examination of the larynx (voice box), February 2005. http://www.pitt.edu/˜ crosen/voice/anatomy2.html.

[26] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on communications*, 28(1):84–95, January 1980.

[27] J. D. Markel. The sift algorithm for fundamental frequency estimation. *IEEE transaction on audio and electroacoustics*, AU-20(5):367–377, December 1972.

[28] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation. an overview. *Digital Signal Processing*, 10:1–18, 2000.

[29] National institute of standars and technology. www page, June 2005. http://www.nist.gov/speech/tests/spk/1999/euro99_v2/index.htm.

[30] Noise data. www page, June 2005. http://spib.rice.edu/spib/select_noise.html.

[31] F. Nolan. *The phonetic bases of Speaker recognition.* Cambridge: Cambridge University Press, 1983.

[32] A. V. Oppenheim and R. W. Schafer. From frequency to quefrency: A history of cepstrum. *IEEE signal processing Magazine*, September 2004.

[33] Principal organs of articulation. www page, December 2004. http://www.sil.org/mexico/ling/glosario/E005bi-OrgansArt.htm.

[34] Praat: doing phonetics by computer. www page, December 2003. http://www.praat.org/.

[35] W. H. Press, B. P. Flannery, and S. A. Teukolsky. *Numerical recipes in C.* Cambridge University Press, 1992.

[36] R. P. Ramachandran, K. R. Farrel, R. Ramachandran, and R. J. Mammone. Speaker recognition - general classifier approaches and data fusion methods. *Pattern recognition society*, 35:2801 – 2821, 2002.

[37] D. A. Reynolds. An overview of automatic speaker recognition technology. In *Proceedings of International conference on acoustics, speech and signal processing (ICASSP)*, volume 4, pages IV–4072 – IV–4075, May 2002.

[38] P. Rose. *Forensic Speaker Identification.* Forensic Science Series. Taylor & Francis, London and New York, 2002.

[39] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley. Average magnitude difference function pitch extractor. *IEEE transactions on acoustics, speech and signal processing*, ASS-22(5):353–362, October 1974.

[40] P. Rubin and E. Vatikiotis-Bateson. *Animal Acoustic Communication*, chapter 8: Measuring and modeling speech production. Springer-Verlag, Berlin/Heidelberg, 1998.

[41] J. Saastamoinen, E. Karpov, V. Hautamäki, and P. Fränti. Automatic speaker recognition for series 60 mobile devices. In *9th International Conference Speech and Computer (SPECOM)*, pages 353 – 360, St. Petersburg, Russia, September 2004.

[42] J. Saastamoinen, E. Karpov, V. Hautamäki, and P. Fränti. Accuracy of MFCC based speaker recognition in series 60 device. *Journal on Applied Signal Processing (EURASIP)*, 2005. accepted.

[43] M. K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg. A lognormal tied mixture model of pitch for prosody based speaker recognition. In *Proceedings of 5th European conference on speech, communication and technology (EUROSPEECH)*, pages 1391–1394, Rhodes, Greece, September 1997.

[44] I. R. Titze. *Principles of voice production.* Prentice Hall, 1994.

[45] E. W. Weisstein. “correlation.” from mathworld–a wolfram. www page, February 2005. http://mathworld.wolfram.com/Correlation.html.

[46] E. W. Weisstein. “periodic function.” from mathworld–a wolfram. www page, February 2005. http://mathworld.wolfram.com/PeriodicFunction.html.

[47] G. S. Ying, L. H. Jamieson, , and C. D. Mitchell. A probabilistic approach to amdf pitch detection. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 1201 –1204, Philadelphia, PA, USA, October 1996.

Appendix A

Tuning set

Table A.1: Filenames for parameter tuning set

Training
1106a.wav 1231a.wav 1831a.wav 3241a.wav 3297a.wav
3764a.wav 4011a.wav 4016a.wav 4018a.wav 4030a.wav
4036a.wav 4041a.wav 4045a.wav 4047a.wav 4049a.wav
4059a.wav 4060a.wav 4063a.wav 4072a.wav 4074a.wav
4081a.wav 4101a.wav 4104a.wav 4105a.wav 4107a.wav
4108a.wav 4110a.wav 4113a.wav 4119a.wav 4124a.wav
4129a.wav 4134a.wav 4137a.wav 4143a.wav 4145a.wav
4148a.wav 4149a.wav 4150a.wav 4154a.wav 4156a.wav
4160a.wav 4162a.wav 4173a.wav 4179a.wav 4184a.wav
4187a.wav 4192a.wav 4194a.wav 4206a.wav 4213a.wav
Testing
1106b.wav 1231b.wav 1831b.wav 3241b.wav 3297b.wav
3764b.wav 4011b.wav 4016b.wav 4018b.wav 4030b.wav
4036b.wav 4041b.wav 4045b.wav 4047b.wav 4049b.wav
4059b.wav 4060b.wav 4063b.wav 4072b.wav 4074b.wav
4081b.wav 4101b.wav 4104b.wav 4105b.wav 4107b.wav
4108b.wav 4110b.wav 4113b.wav 4119b.wav 4124b.wav
4129b.wav 4134b.wav 4137b.wav 4143b.wav 4145b.wav
4148b.wav 4149b.wav 4150b.wav 4154b.wav 4156b.wav
4160b.wav 4162b.wav 4173b.wav 4179b.wav 4184b.wav
4187b.wav 4192b.wav 4194b.wav 4206b.wav 4213b.wav