Universidade Federal de Pernambuco

Centro de Informática

A Fractional Gaussian Mixture Model for Speaker Verification

Final Term Paper

Eduardo Martins Barros de Albuquerque Tenório

March 3, 2015

Declaration

This paper is a presentation of my original research work, as partial fulfillment of the requirement for the degree in Computer Engineering. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

The work was done under the guidance of Prof. Dr. Tsang Ing Ren, at Centro de Informática, Universidade Federal de Pernambuco, Brazil.

Eduardo Martins Barros de Albuquerque Tenório

In my capacity as supervisor of the candidate's paper, I certify that the above statements are true to the best of my knowledge.

Prof. Dr. Tsang Ing Ren

March 3, 2015

Acknowledgements

I am thankful to my parents, for the support and patience during the graduation,
To my adviser, Tsang Ing Ren, for the guidance,
To Cleice Souza, for the previous readings and help.

Live long and prosper

Vulcan salute

Abstract

TODO escrever o abstract após terminar tudo (após a conclusão)

Contents

1. Introduction

The rise in popularity and naturality of computational systems in the everyday of modern life creates the need for easy and less invasive forms of authentication. While enter a hard to memorize password in a terminal still is the safest approach, voice biometrics presents itself as a continuing improvement alternative. Also, speech is the most natural way humans have to communicate, being incredibly complex and with numerous specific details related to the speaker [1]. Therefore, it is expected an increasing usage of vocal interfaces to perform actions such as computer login, voice search (e.g., Apple Siri, Google Now and Samsung S Voice) and identification of speakers in a conversation and its content.

At present, fingerprint biometrics is adopted in several solutions (e.g., ATMs [2]), authentication through facial recognition comes as built-in software for average computers and iris scan was adopted for a short time by United Kingdom and permanently by United Arab Emirates border controls [3, 4]. That said, improvements in voice recognition techniques indicate a near future where vocal commands will be used for authentication, alone or combined with other biometric methods.

Current commercial products based on voice technology (e.g., Dragon Naturally Speaking, KIVOX and VeriSpeak) are usually intended to perform either **speech recognition** (*what* is being said) or **speaker recognition** (*who* is speaking). Voice search applications are designed to determine the content of a speech, with no concern about who the speaker is or if there is more than one, while computer login and telephone fraud prevention supplement a memorized personal identification code with speaker verification [5]. Few applications need to perform both processes, such as automatic speaker labeling of recorded meetings, that transcribes what each person is saying. To achieve this goal, numerous voice processing techniques have become known in industry and academy, e.g., Natural Language Processing (NLP), Hidden Markov Models (HMM) and GMMs. Although all of these are state-of-the-art, the subject covered by this paper is a subarea of speaker recognition and only a small subset of techniques will be unraveled.

1.1 Speaker Recognition

As stated in [6], speaker recognition may be divided in two subareas. The first is **speaker identification**, aimed to determine the identity of a speaker (through a speech signal) from a non-unitary set of known speakers. This task is also named speaker identification in **closed set**. In the second, **speaker verification**, the goal is to determine if a speaker is who he or she claims to be, not an imposter. As the set of imposters is unknown *a priori*, this is an **open set** problem. An intermediate task is **speaker identification in open set**, when an "imposter class" is added in order to categorize all unmatched speakers.

The text used may have constraints, such as by class (e.g., digits and language) and/or number of words used (e.g., one word or sentences). In **text-dependent** systems the content of the speech is relevant to the evaluation, and the training and test utterances must contain the same spoken text, e.g. a passphrase. A change in the text used demands an entirely new training section. **Text-independent** systems have no restrictions to the message in both sets, with the non-textual characteristics of the user's voice (e.g., pitch and accent) being the important aspects to the evaluator. These characteristics are presented in different sentences, usage of different languages and even in gibberish for a speaker. Between the extremes in constraints falls the **vocabulary-dependent system**, which constrains the speech to come from a limited vocabulary (e.g., digits) from which test words or phrases are selected (e.g., "one-two-three") [5].

This paper is focused in **text-independent speaker verification**, in other words, the acceptance or rejection of a user's claimed identity by analysis of his or her vocal characteristics with no specific text. To achieve that, a speaker's GMM adapted from an UBM [7] is implemented. Also, an adaptation of the technique is proposed and evaluated using the theory of FCM presented in [8].

1.2 Objectives

The objectives of this study are:

- Analyze and evaluate the speaker verification system using the adapted GMM discussed in [7];
- Propose and evaluate a new method derived from GMM, using the FCM theory discussed in [8];
- Conduct experiments for the existent and the proposed methods and perform comparisons.

1.3 Document Structure

Chapter 2 contains basic information about voice recognition, as well as the basic architecture for a speaker verification system. The feature extraction process is explained in chapter 3, from the reasons for its use to the chosen technique (MFCC). In chapter 4 is detailed the GMM and the UBM-GMM. Chapter 5 introduces FCM and the proposed FGMM. Experiments are described in chapter 6, as well as its results. Finally, chapter 7 concludes the study. Furthermore, this work contains an appendix with the most relevant pieces of the source code and some necessary mathematical concepts.

2. Speaker Recognition Systems

The process of voice recognition lies on the field of pattern classification, with the speaker and his or her utterance (a speech signal) as inputs for a classifier and a decision as output. This decision may be, given an utterance $\boldsymbol{Y}$ produced by a speaker $\mathcal{S}$ and a set $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, ..., \mathcal{S}_S\}$ of known users,

$$\mathcal{S} \leftarrow \mathcal{S}_i, \text{ if } i = \arg\max_j P(\mathcal{S}_j|\boldsymbol{Y}). \tag{2.1}$$

This is a case of speaker identification and the output is a $\mathcal{S}_i$ from $\boldsymbol{\mathcal{S}}$. Another type of decision is

$$\text{if } P(\mathcal{S}_i|\boldsymbol{Y}) \begin{cases} \geq \alpha, & \text{accept } \mathcal{S}, \\ < \alpha, & \text{reject } \mathcal{S}. \end{cases} \tag{2.2}$$

This is a speaker verification decision, with a binary output, given a $\mathcal{S}$ who produced $\boldsymbol{Y}$, a claimed identity $\mathcal{S}_i$ from $\boldsymbol{\mathcal{S}}$ and a threshold α for acceptance. This chapter (and indirectly the whole document) is about the type of decision from Eq. 2.2.

2.1 Basic Concepts

2.1.1 Utterance

An utterance is a piece of speech produced by a speaker. It may be a word, a statement or any vocal sound. The terms *utterance* and *speech signal* sometimes are used interchangeably, but here speech signal will be associated to an utterance recorded and digitalized. An example of an utterance as speech signal is shown in Fig. 2.1.

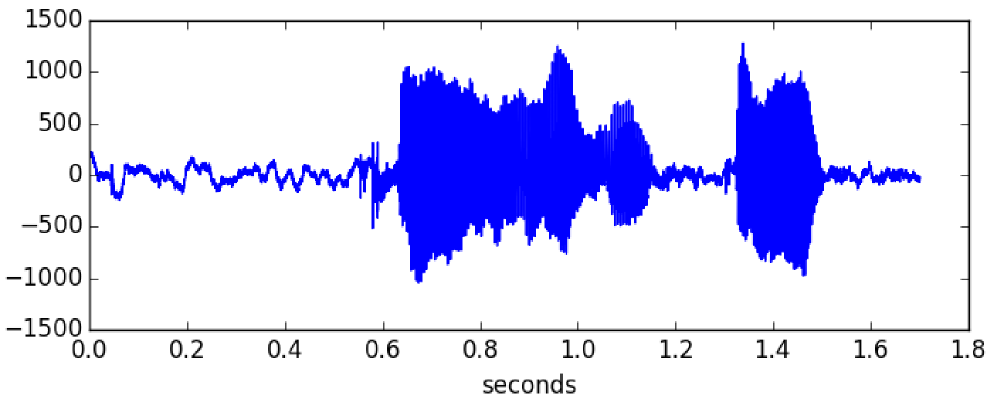

Figure 2.1: Speech signal for utterance “karen livescu”.

2.1.2 Features

The raw speech signal is unfit for usage by a recognition system. For a correct processing, the unique features from the speaker's vocal tract are extracted, reducing the number of variables the system needs to deal with (leading to a simpler implementation) and performing a better evaluation (just use the necessary informations). This extraction is executed by the MFCC algorithm, explained in Chapter 3. Due to the stationary properties of the speech signal when analyzed in a short period of time, it is divided in overlapping frames of small and predefined length, to avoid "loss of significancy" in the features [9, 10].

2.1.3 Dependency x Independency

2.2 Likelihood Ratio Test

TODO basear-se na seção 2 do artigo "Speaker Verification Using Adapted Gaussian Mixture Models".

2.3 Basic Speaker Verification Architecture

2.3.1 Training Phase

2.3.2 Test Phase

3. Feature Extraction

As an acoustic wave propagated through space over time, the speech signal is not appropriate to be evaluated by the speaker verification system. In order to deliver decent outcomes, a good parametric representation must be provided to the system. This task is performed by the feature extraction process, which transforms a speech signal into a sequence of characterized measurements, i.e. features. As stated in [9], "the usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data, and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences". According to [11] the ideal features should:

- occur naturally and frequently in normal speech;
- be easily measurable;
- vary highly among speakers and be very consistent for each speaker;
- not change over time nor be affected by the speaker's health;
- be robust to reasonable background noise and to transmission characteristics;
- be difficult to be artificially produced;
- not be easily modifiable by the speaker.

Features may be categorized based on vocal tract or behavioral aspects, divided in (1) short-time spectral, (2) spectro-temporal, (3) prosodic and (4) high level [6]. Short-time spectral features are usually calculated using millisecond length windows and describe the voice spectral envelope, composed of supralaryngeal properties of the vocal tract, e.g. timbre. Prosodic and spectro-temporal occur over time, e.g. rhythm and intonation, and high level features occur during the conversation, e.g. accents.

The parametric representations evaluated in [9] may be divided into those based on the Fourier spectrum, Mel-Frequency Cepstrum Coefficients (MFCC) and Linear Frequency Cepstrum Coefficients (LFCC), and those based on the Linear Prediction Spectrum, Linear Prediction Coefficients (LPC), Reflection Coefficients (RC) and Linear Prediction Cepstrum Coefficients (LPCC). The better evaluated representation was the MFCC, with minimum and maximum accuracy of 90.2% and 99.4% respectively, leading to its choice as the parametric representation in this work.

3.1 Mel-Frequency Cepstral Coefficient

MFCC is a highly used parametric representation in the area of voice processing, due to its similarity with the mode the human ear operates. Despite the fact the ear is divided in three sections, i.e. outer, middle and inner ears, only the last is mimicked. The mechanical pressure waves produced by the triad hammer-anvil-stirrup are received by the cochlea (Fig. 3.1), a spiral-shaped cavity with a set of inner hair cells attached to a membrane (the basilar membrane) and filled with a liquid. This structure converts motion to neural activity through a non-uniform spectral analysis [10] and passes it to the pattern recognition in the brain.

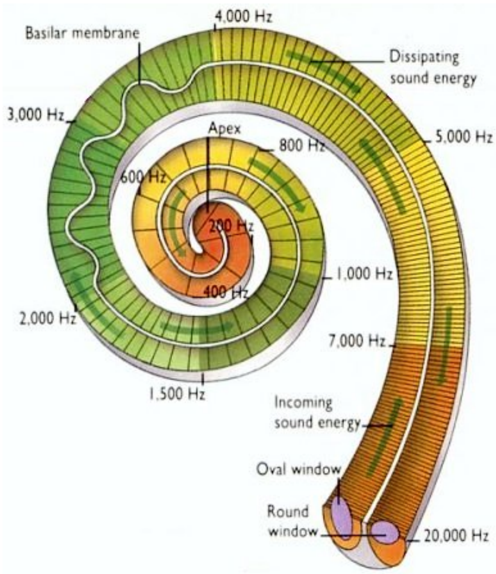

Figure 3.1: Cochlea divided by frequency regions.

A key factor in the perception of speech and other sounds is *loudness*, a quality related to the physical property of sound pressure level. Loudness is quantified by relating the actual sound pressure level of a pure tone (in dB relative to a standard reference level) to the perceived loudness of the same tone (in a unit called phons) over the range of human hearing (20 Hz–20 kHz) [10]. As shown in Fig. 3.2, a 100 Hz tone at 60 dB is equal in loudness to a 1000 Hz tone at 50 dB, both having the *loudness level* of 50 phons (by convention).

3.1.1 The Mel Scale

The mel scale is the result of an experiment conducted by Stevens, Volkmann and Newman [13] intended to measure the perception of a pitch and construct a scale based on it. Each observer was asked to listen to two tones, one in the fixed frequencies 125, 200, 300, 400, 700, 1000, 2000, 5000, 8000 and 12000 Hz, and the other free to have its frequency varied by the observer for each fixed frequency of the first tone. An interval of 2 seconds separated both tones. The observers were instructed to say in which frequency the second tone was "half the loudness" of the first. A geometric mean was taken from the observers' answers and a measure of 1000 mels was assigned to the frequency of 1000 Hz, 500 mels to the frequency sounding half as high (as determined by Fig. 1 in [13]) and so on.

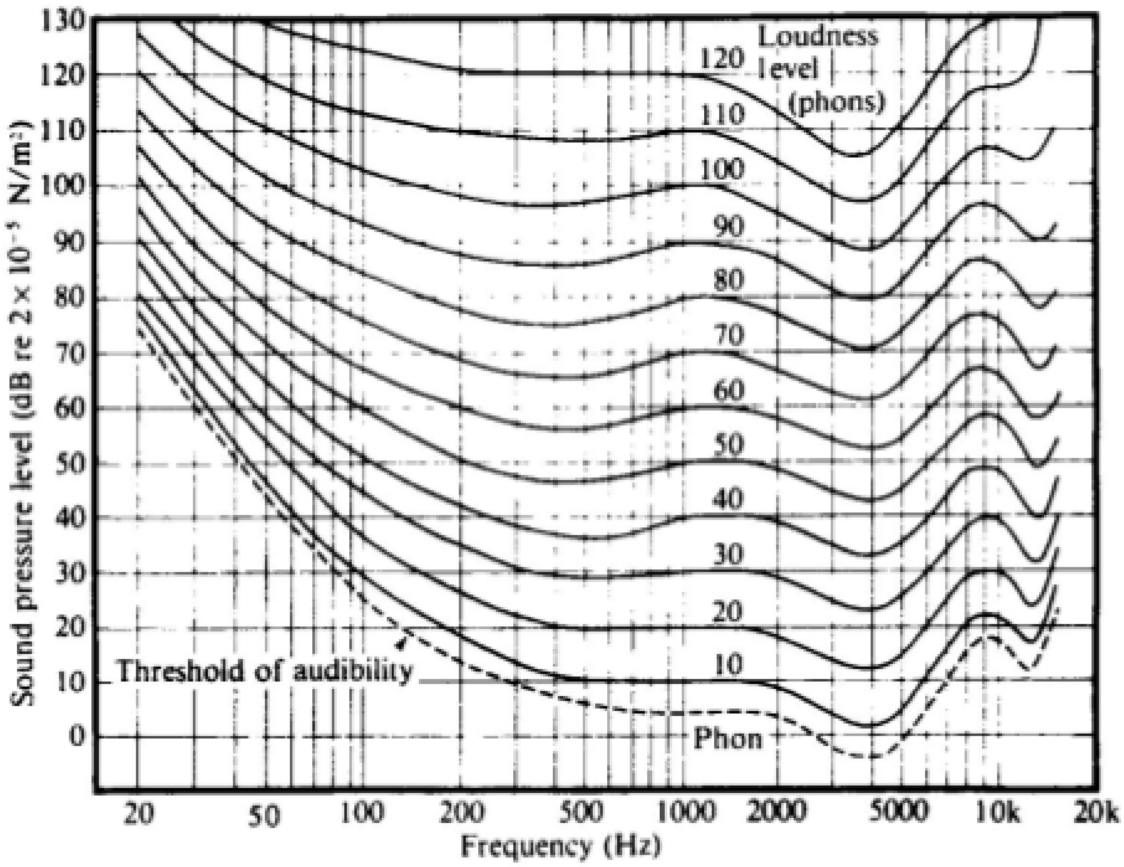

Figure 3.2: Loudness level for human hearing [12].

Decades after the creation of the mel scale, O'Shaughnessy [14] published an equation to convert frequencies in hertz to frequencies in mels.

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \tag{3.1}$$

Being logarithmic, the growth of a mel-frequency curve is slow with a linear growth of the frequency in hertz. Eq. 3.1 sometimes is used only for frequencies higher than 1000 Hz while the lower frequencies obey a linear function. In this work all conversions will use Eq. 3.1, as shown by Fig. 3.3.

3.1.2 Cepstrum

3.1.3 Extraction Process

Pre-emphasis

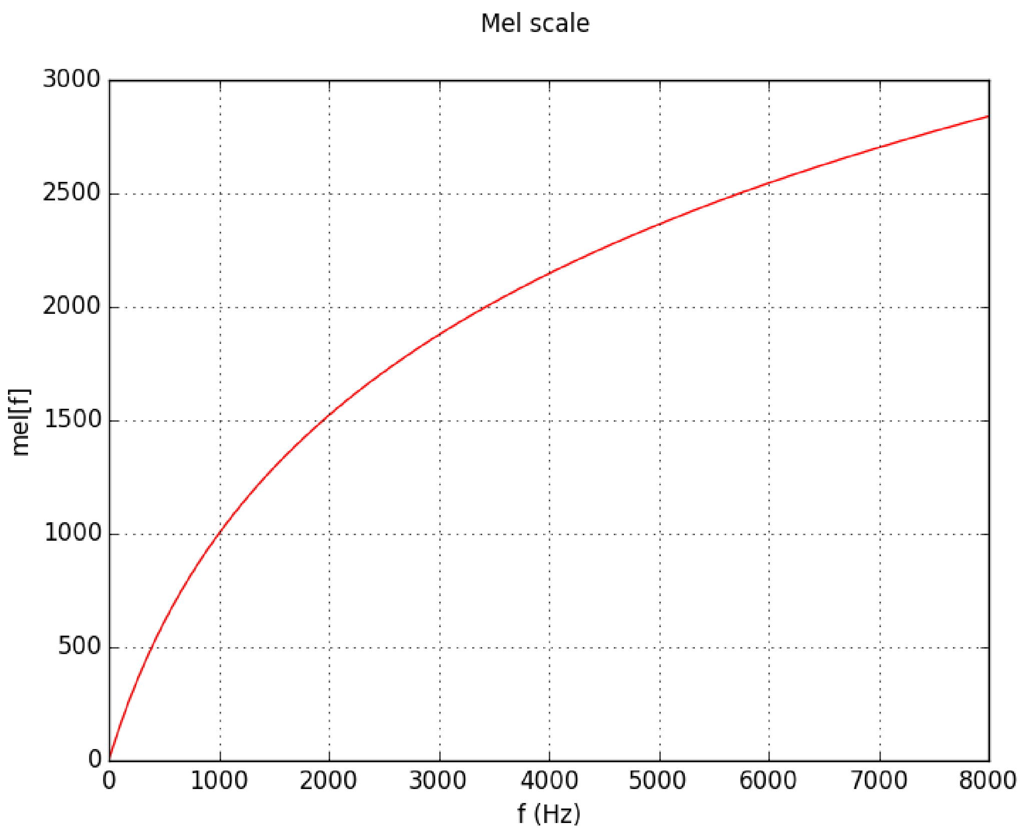

Figure 3.3: The logarithm curve of the mel-frequency.

4. Gaussian Mixture Model

5. Fractional Gaussian Mixture Model

6. Experiments

7. Conclusion

TODO escrever a conclusão após terminar tudo (antes do abstract)

A. Codes

References

[1] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz and Douglas A. Reynolds. “A Tutorial on Text-Independent Speaker Verification”. In: *EURASIP Journal on Applied Signal Processing* 4 (2004), pp. 430–451.

[2] P.T. Wang and S.M. Wu. “Personal fingerprint authentication method of bank card and credit card”. Pat. US Patent App. 09/849,279. 2002. URL: `https://www.google.com/patents/US20020163421`.

[3] M. Angela Sasse. “Red-Eye Blink, Bendy Shuffle, and the Yuck Factor: A User Experience of Biometric Airport Systems”. In: *Security & Privacy, IEEE* 5.3 (2007), pp. 78–81.

[4] Ahmad N. Al-Raisi and Ali M. Al-Khouri. “Iris recognition and the challenge of homeland and border control security in UAE”. In: *Telematics and Informatics* 25.2 (2008), pp. 117–132.

[5] Douglas A. Reynolds. “Automatic Speaker Recognition Using Gaussian Mixture Speaker Models”. In: *The Lincoln Laboratory Journal* 8.2 (1995), pp. 173–192.

[6] Hector N. B. Pinheiro. *Sistemas de Reconhecimento de Locutor Independente de Texto*. Trabalho de Graduação. Universidade Federal de Pernambuco, 2013.

[7] Douglas A. Reynolds, Thomas F. Quatieri and Robert B. Dunn. “Speaker Verification Using Adapted Gaussian Mixture Models”. In: *Digital Signal Processing* 10.1 (2000), pp. 19–41.

[8] Chaobang Gao, Jiliu Zhou and Qiang Pu. “Theory of fractional covariance matrix and its applications in PCA and 2D-PCA”. In: *Expert Systems with Applications* 40.13 (2013), 5395–5401.

[9] Steven B. Davis and Paul Mermelstein. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-28.4 (1980), pp. 357–366.

[10] Lawrence R. Rabiner and Ronald W. Schafer. “Introduction to Digital Speech Processing”. In: *Foundations and Trends in Signal Processing* 1.1-2 (2007), pp. 1–194.

[11] Jared J. Wolf. “Efficient acoustic parameters for speaker recognition”. In: *Journal of the Acoustical Society of America* 51 (1972), pp. 2044–2056.

[12] Harvey Fletcher and Wilden A. Munson. “Loudness, Its Definition, Measurement and Calculation”. In: *Bell Telephone Laboratories* 12.4 (1933), pp. 82–108.

[13] Stanley S. Stevens, John Volkmann and Edwin B. Newman. “A Scale for the Measurement of the Psychological Magnitude Pitch”. In: *The Journal of Acoustical Society of America* 8.3 (1937), pp. 185–190.

[14] Douglas O’Shaughnessy. *Speech Communications: Human and Machine*. Addison-Wesley, 1987.