Universidade Federal de Pernambuco

Centro de Informática

Fractional Gaussian Mixture Models for Speaker Verification

Eduardo Martins Barros de Albuquerque Tenório

February 8, 2015

Abstract

TODO EDITAR Abstract goes here

Dedication

TODO EDITAR To mum and dad

Declaration

TODO EDITAR I declare that..

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.

Contents

Chapter 1

Introduction

Chapter 2

Speaker Recognition System

Chapter 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 [1], "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 [2] 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 [3]. 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 [1] 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 [4] and passes it to the pattern recognition in the brain.

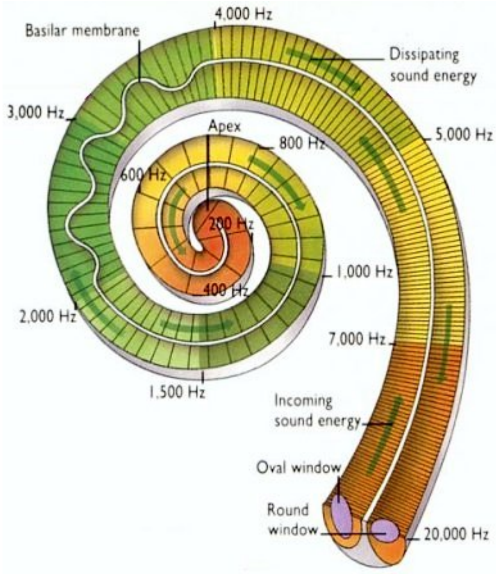

Figure 3.1: Cochlea divided by frequency regions.

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) [4]. 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 [6] 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 [6]) and so on.

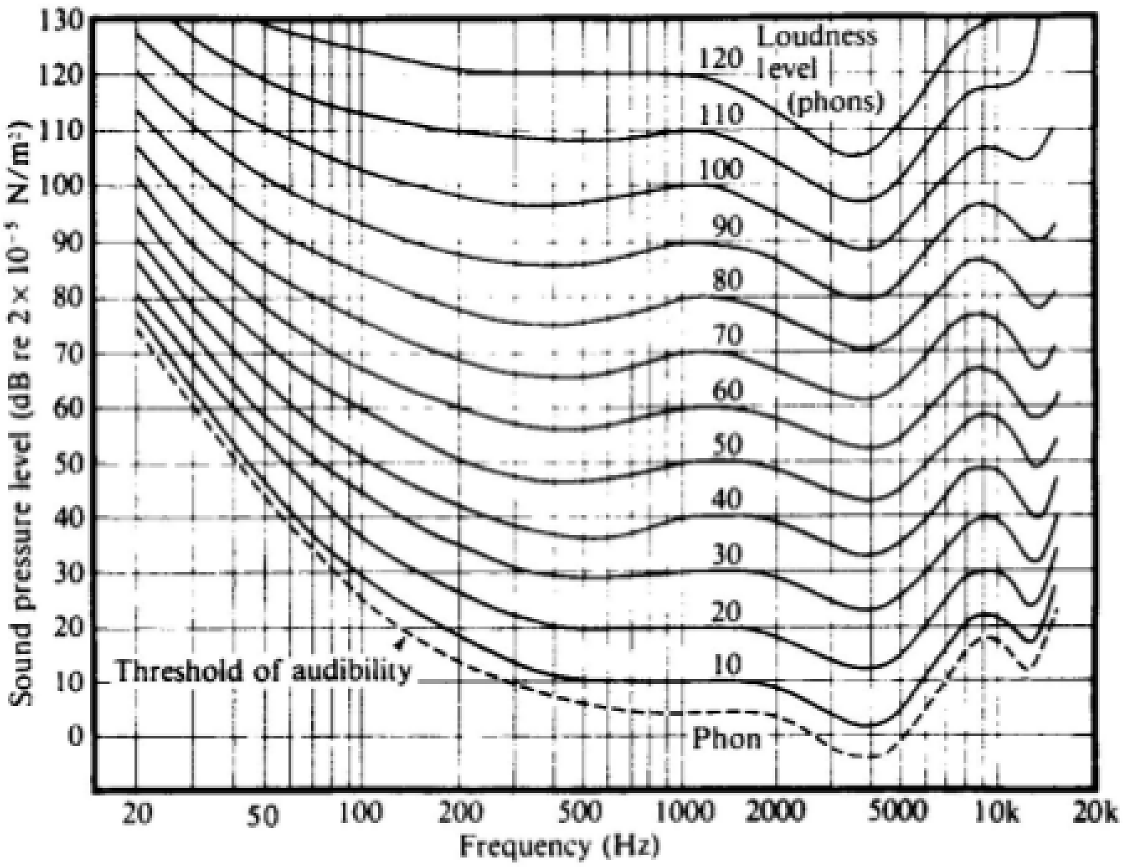

Figure 3.2: Loudness level for human hearing [5].

Decades after the creation of the mel scale, O'Shaughnessy [7] published an equation to convert frequencies in hertz to frequencies in mels.

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

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

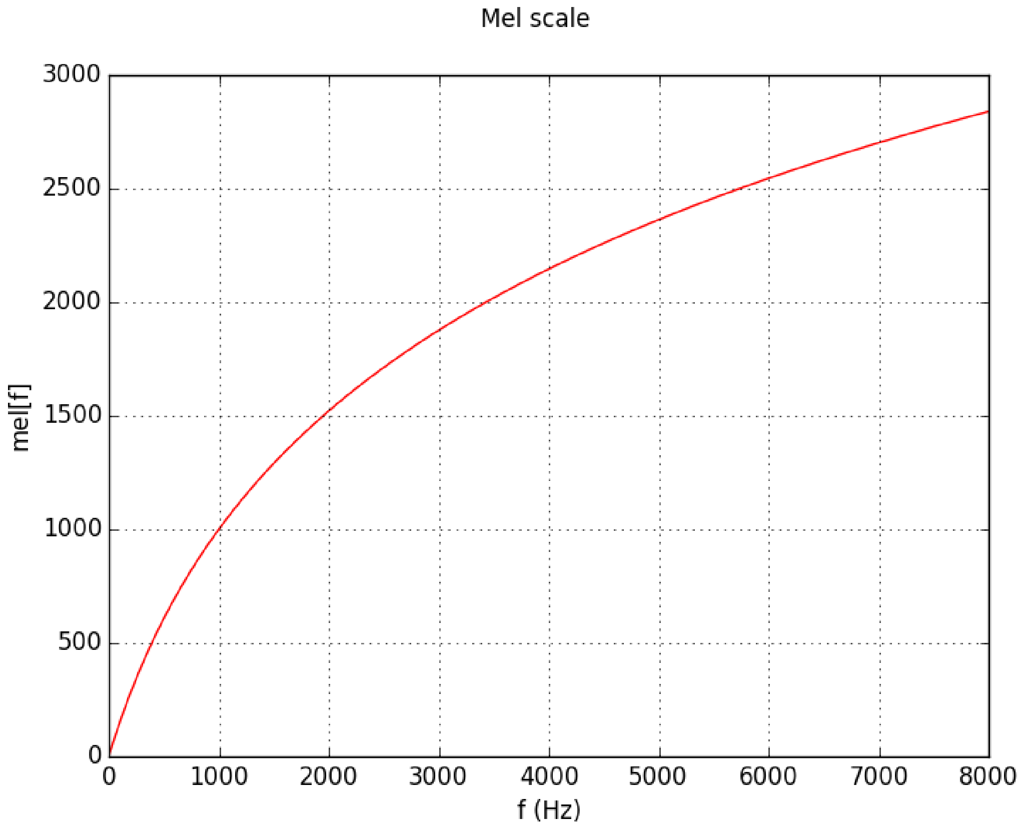

Figure 3.3: The logarithm curve of the mel-frequency.

Chapter 4

Gaussian Mixture Models

Chapter 5

Fractional Covariance Matrix

Chapter 6

Experiments

Chapter 7

Conclusion

Appendix A

Codes

Bibliography

[1] 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.

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

[3] Hector N. B. Pinheiro. *Sistemas de Reconhecimento de Locutor Independente de Texto*. Major Paper. Universidade Federal de Pernambuco, 2013.

[4] 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.

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

[6] J. Volkmann S. S. Stevens and E. 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.

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