Universidade Federal de Pernambuco

Centro de Informática

Improvements in a Gaussian Mixture Models based Speaker Verification System using Fractional Covariance Matrix

Eduardo Martins Barros de Albuquerque Tenório

January 16, 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 (features). As stated in [1], the usual objectives in selecting a representation are (1) to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data, and (2) 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 as much as possible among speakers, but be as consistent as possible for each speaker;
- not change over time or be affected by the speaker's health;
- not be affected by reasonable background noise nor depend on specific transmission characteristics;
- not be modifiable by conscious effort of the speaker, or, at least, be unlikely to be affected by attempts to disguise the voice.

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 usually are 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 the most used parametric representation in the area of voice processing, due to its similarity with the human ear operation. 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 hammer, the anvil and the stirrup is received by the cochlea (Fig. 3.1), a spiral-shaped cavity with a set of inner hair cells (basilar membrane) that converts motion to neural activity throught a non-uniform spectral analysis [4]. This activity is then passed to the pattern recognition existent in the brain.

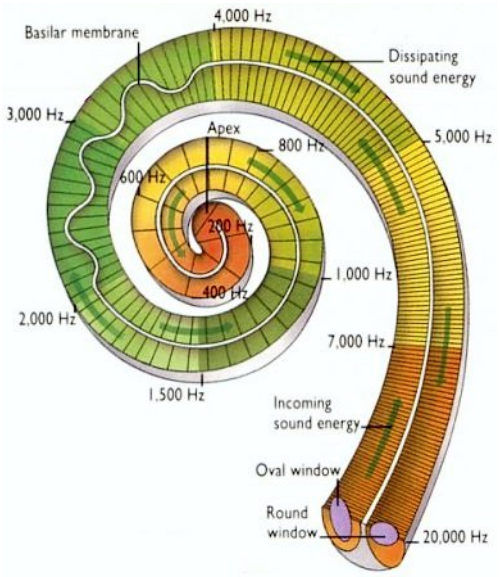

Figure 3.1: Cochlea divided by frequency regions.

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]. This relationship is shown in Fig. 3.2.

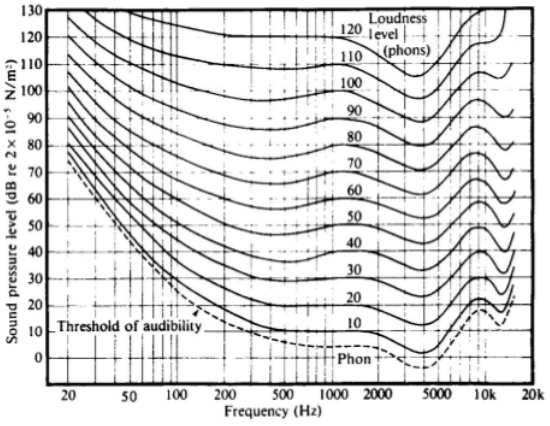

Figure 3.2: Loudness level for human hearing.

3.1.1 The Mel Scale

3.1.2 Cepstrum

3.1.3 Extraction Process

Pre-emphasis

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.