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 15, 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 used by the speaker verification system. In order to deliver better 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). 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 [1]. 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 overtime 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 parametric 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

TODO explicar brevemente o MFCC e quais suas vantagens

3.1.1 The Mel Scale

3.1.2 Extraction Process

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.