JFA based Speaker Recognition using Delta-Phase and MFCC features

Ahilan Kanagasundaram, David Dean, Sridha Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology, Brisbane, Australia
{a.kanagasundaram, d.dean, s.sridharan }@qut.edu.au

Abstract

This paper investigates the use of mel-frequency delta-phase (MFDP) features in comparison to, and in fusion with, traditional mel-frequency cepstral coefficient (MFCC) features within joint factor analysis (JFA) speaker verification. MFCC features, commonly used in speaker recognition systems, are derived purely from the magnitude spectrum, with the phase spectrum completely discarded. In this paper, we investigate if features derived from the phase spectrum can provide additional speaker discriminant information to the traditional MFCC approach in a JFA based speaker verification system. Results are presented which provide a comparison of MFCC-only, MFDP-only and score fusion of the two approaches within a JFA speaker verification approach. Based upon the results presented using the NIST 2008 Speaker Recognition Evaluation (SRE) dataset, we believe that, while MFDP features alone cannot compete with MFCC features, MFDP can provide complementary information that result in improved speaker verification performance when both approaches are combined in score fusion, particularly in the case of shorter utterances.

Index Terms: speaker verification, MFCC features, JFA, Delta-phase

1. Introduction

In recent speaker verification research, the joint factor analysis (JFA) technique has become one of more successful approaches to speaker verification by explicitly modelling enrolment and verification mismatch. This approach is typically based upon acoustic features derived from the magnitude spectrum, with most approaches using mel-frequency cepstral coefficients (MFCC) to represent the acoustic domain for modelling against the universal background model (UBM) in forming the speaker and channel factors. While there have been investigations of phase-based features for speaker verification using simple Gaussian mixture model (GMM) [1] and support vector machine (SVM) [2] approaches, no investigation has yet been performed using phase-based features in the explicit channel and speaker modelling approach taken by JFA speaker verification systems.

In order to make use of the phase spectrum for speaker verification, it needs to be transformed into a meaningful representation that provides adequate discrimination between individual speakers. One of the first attempts at using phase-based features for automatic speaker recognition was through the use of modified group delay function (GDF) by Murthy *et al.* [3], defined as the frequency-domain derivative of the phase spectrum, modified to attenuate the effect of zeros in the z-plane of the frequency representation. This approach was shown to outperform MFCC speaker verification using a GMM-UBM modelling approach. [3].

An alternative approach to constructing phase-based features was introduced by Wang *et al.* by looking at the time-domain derivative of the phase spectrum, termed the instantaneous frequency deviation (IFD) [4]. This work was further extended by McCowan *et al.* to develop the mel-frequency delta-phase (MFDP) representation [2] and demonstrated its performance to be similar to that of MFCC using a modern GMM-supervector-based SVM approach, but only without channel compensation. When feature warping and nuisance attribute projection (NAP) were applied to both the MFCC and MFDP systems, the MFDP system was found lacking in comparison to the MFCC. However, even though the channel-compensated MFDP system was not comparable to the MFCC approach individually it was still shown to provide complementary information in fusion with the MFCC, with a score fusion approach outperforming both individual approaches.

In this paper, we study the use of MFDP features introduced by McCowan *et al.* [2], in a modern JFA-based speaker verification system in order to investigate the ability of the explicit speaker and channel modelling to cope with phase-based features. Initially both MFDP and MFCC features will be studied individually within a JFA speaker verification system with a combination of experimental parameters to determine the best individual approach for both sets of features. Thereafter the best configuration of MFDP and MFCC features will be combined to analyze the fused JFA system.

Throughout this paper, both medium length and short utterances will be evaluated to determine if the performance of MFDP features vary according to the amounts of speech available for enrolment and verification. This approach has been taken before for MFCC features in both JFA [5], SVM [6], and i-vector [7] speaker verification systems but no similar studies have been performed on phase-based features.

2. Mel-frequency delta-phase features

The process of extracting MFDP features from the acoustic speech is designed to attempt extract speech information from the phase-domain through calculating a phase difference between successive frames separated by a short time interval [2]. The delta-phase spectrum can be calculated from the Fourier transform of two successive frames $\tilde{X}_m(k)$ and $\tilde{X}_{m-1}(k)$, as follows:

$$\Delta\phi_m(k) = arg\left[\frac{\tilde{X}_m(k)e^{-j\omega_k mD}}{\tilde{X}_{m-1}(k)e^{-j\omega_k (m-1)D}}\right] \tag{1}$$

$$|\Delta\phi_m(k)| = |arg\left[\left(\frac{\tilde{X}_m(k)}{\tilde{X}_{m-1}(k)}\right)e^{-j\omega_k D}\right]| \tag{2}$$