Group Delay Features for Speaker Recognition

Tharmarajah Thiruvaran[1,2], Eliathamby Ambikairajah[1,2], Julien Epps[1]
[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052 Australia.
[2]National Information Communication Technology (NICTA),
Australian Technology Park, Eveleigh 1430, Australia.
thiruvaran@student.unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

***Abstract*— Group delay is proposed as an effective means of representing spectral phase information as a feature in speaker recognition. Robustness of group delay features is difficult to achieve, since the spiky nature of the group delay masks the fine structure of the group delay. In this paper, two features based on group delay are proposed by reducing the effect of spikes with two different approaches. The first is log compression, to address the masking effects of the spikes, and the second is to use a sub-band based approach, where masking is restricted within certain bands containing the spikes. The purpose of this paper is to introduce different types of group delay feature extraction methods. The two features are evaluated on the cellular NIST 2001 database.**

***Keywords*—speaker recognition, group delay**

I. INTRODUCTION

Speaker recognition is the problem of automatically recognizing speakers using only their speech signal. Conventionally, the front-end of the recognition system uses amplitude based features such as Mel frequency cepstral coefficients (MFCC). As the performance with this front-end alone is not sufficient for real world industrial applications, recently phase information from the speech signal has been explored [1-3]. Previously, phase information has received less attention due to the difficulties in representing it as a feature. This is mainly because the phase distribution is wrapped within $\pm\pi$ which requires unwrapping [1]. However, the variation of phase can be utilized for speech processing with meaningful interpretations such as the group delay [2], which can be calculated directly from the signal without using any phase unwrapping.

The group delay of speech suffers from spiky characteristic [2], resulting in difficulties for use in speech processing applications. The spikes are mainly due to the excitation, and mask the effect of vocal tract resonances in the group delay. A good illustration of this masking effect is given in [2]. In order to suppress the spikes various modifications have been proposed [2, 4, 5]. Variants on group delay features were used in several speech processing applications such as phoneme recognition [5, 6], speech recognition [7, 8], language identification [6, 9], speaker identification [6, 10, 11], glottal flow estimation [12] and formant extraction [13].

To suppress peaks in the group delay, a major modification [2, 5] has been to replace the power spectrum with the cepstrally smoothed version of the power spectrum in the group delay calculation. However, this modification alone is not enough to remove the masking effects of the spikes. To improve further, other modifications have been proposed. One modification is based on two empirical parameters to further suppress the peaks [5]. As stated in [5] the lack of theoretical insight for the parameters requires an empirical search of all possible combinations, to select the optimum combination for that particular task. This is very time consuming in experiments with large database and is not desirable for real world applications where data independent algorithms are preferred. The other method [4] removes all but the slow variations in the group delay by passing it through a low pass filter. Here, the low pass filtering not only removes the peaks but also removes the fine structure of the group delay and the performance depends on the cut off frequency of the filter.

In an attempt to circumvent these problems, we propose two alternative group delay features. One uses log compression to suppress the effects of peaks while retaining the fine structure, and the other introduces a sub-band approach to group delay extraction. These two features are evaluated on the cellular NIST 2001 speaker recognition database.

II. GROUP DELAY OF SPEECH

A. Extraction of Group Delay from speech

A brief summary of the methods used to extract group delay in speech is provided in this section. Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. It is the time-domain delay of each frequency component of the signal, as a function of frequency. For a (continuous-time) signal $x(t)$, its Fourier transform $X(f)$ can be written as

$$X(f) = \left|X(f)\right| \exp^{j\phi(f)}. \tag{1}$$

Then the group delay is defined as in (2).

$$G(f) = -\frac{d\phi(f)}{df} \tag{2}$$

In implementation, the phase should be unwrapped before taking the derivative, which is problematic [1]. To avoid unwrapping, another method [4] calculates the group delay using only amplitude values, as in (3).

1-4244-0983-7/07/$25.00 ©2007 IEEE