LS Regularization of Group Delay Features for Speaker Recognition

Jia Min Karen Kua1,2, Julien Epps1,2, Eliathamby Ambikairajah1,2, Eric Choi2

1School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia
2ATP Research Laboratory, National ICT Australia (NICTA), Evelyne 2015, Australia

Abstract

Due to the increasing use of fusion in speaker recognition systems, features that are complementary to MFCCs offer opportunities to advance the state of the art. One promising feature is based on group delay, however this can suffer large variability due to its numerical formulation. In this paper, we investigate reducing this variability in group delay features with least squares regularization. Evaluations on the NIST 2001 and 2008 SRE databases show a relative improvement of at least 6% and 18% EER respectively when group delay-based system is fused with MFCC-based system.

Index Terms: speaker recognition, group delay, least squares regularization

1. Introduction

Speaker recognition is the problem of determining a person’s identity based on the intrinsic characteristics of his/her voice. A widely used speech production model is the source-filter model. In the source-filter model, the vocal tract system is excited by the vocal source excitation signal, which is acoustically equivalent to the glottal airflow originated from the lungs and modulated at the larynx [1]. The most representative vocal tract-related acoustic feature is the Mel-frequency cepstral coefficients (MFCC), whose success has seen it become a de facto standard feature for speaker recognition. The usefulness of vocal source-related features, on the other hand, has also been investigated, though to a lesser extent. Such features include mainly pitch, harmonic structure, and phase information [2]. Research reported in [3] and [4] has attempted to estimate and model the glottal flow derivative waveform and used these parameters to identify individual speaker. These studies have demonstrated that the vocal source-related features provide complementary information to MFCC. After a preliminary empirical study of different features, phase based features, in particular derived frequency modulation (FM) and group delay representations, showed promise.

Previously, phase based features have seldom been used because the phase spectrum of a signal is wrapped within ±π, requiring unwrapping. The neglect of phase spectra results in a loss of all information pertaining to changes in the spectral content that occur within the duration of a single frame. Although phase unwrapping techniques have been proposed [5], it is generally preferred to avoid unwrapping. Hence, the group delay can be utilized, which contains the same information as the phase spectrum. Alternatively, frequency modulation feature, motivated by the AM-FM model [6], could be utilized.

The modified group delay function [7] has proven successful in recent speech processing research, however it is not ideal as a feature, since high-amplitude peaks, corresponding to the spectral fine structure, create unwanted variability in the feature distributions. These peaks are caused by zeros of the z-transform of the excitation components of the speech signal. Zeros close to the unit circle in the z-plane produce large-amplitude spikes in the group delay function and mask the group delay information corresponding to the vocal tract system [8]. In order to suppress the peaks, various modifications have been proposed, from replacement of the power spectrum with cepstrally smoothed power spectrum [7, 8], inclusion of two empirical parameters [7], low pass filtering [9], to log compression [10] in the group delay calculation.

In an attempt to circumvent the above problems in this paper, we propose alternative group delay features regularized using a least squares approach. The resulting features are evaluated on the NIST2001 and NIST2008 speaker recognition databases.

2. Group delay feature extraction

2.1. Group delay

Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. For a (continuous-time) signal \( x(t) \), its Fourier transform \( X(f) \) can be written as equation (1)

\[
X(f) = |X(f)| e^{j\phi(f)}
\]

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

\[
G(f) = -\frac{d\phi(f)}{df}
\]

The phase spectrum of a signal is difficult to process because it is available in a wrapped form [8]. To avoid unwrapping, a group delay calculation based only on amplitude values was proposed [9]:

\[
G(f) = \left\{ \frac{X(f) n F\{tx(t)\} + X(f) F\{tx(t)\}}{|X(f)|^2} \right\}
\]

where the subscripts \( R \) and \( I \) denote the real and imaginary parts of the Fourier transform \( F \{ \} \) respectively. However, the group delay function requires the signal to be minimum phase or that the poles of the transfer function be within the unit circle in order to achieve a smooth estimate. The group delay function becomes peaky in nature due to pitch peaks, noise and window effects [7, 11, 12]. It is also important to note that the denominator term \( |X(f)|^2 \) in equation (3) becomes zero, at zeros that are located close to the unit circle. To suppress the peaky characteristics of the group delay, a cepstrally smoothed power spectrum \( |S(f)|^2 \) was proposed to replace \( |X(f)|^2 \) [8]:
Figure 1: Comparison of group delay spectra for a voiced frame of speech. As expected, longer regularization windows $M$ produce smoother spectra.

$$G_s(f) = \left\{ \frac{X(f)R F\{tx(t)\}_R + X(f)\gamma F\{tx(t)\}_I}{|S(f)|^2} \right\}$$  \hspace{1cm} (4)

This equation was further modified to include two new parameters, $\alpha$ and $\gamma$. The resulting group delay is termed the modified group delay (MODGD) [7] as in equation (5):

$$G_{mod}(f) = \text{sign}\left| \frac{X(f)R F\{tx(t)\}_R + X(f)\gamma F\{tx(t)\}_I}{|S(f)|^{2\gamma}} \right|^\alpha$$  \hspace{1cm} (5)

where $\text{sign}$ is the sign of $G_s(f)$ in equation (4). However the values of $\alpha$ and $\gamma \in [0,1]$ have to be determined experimentally [12]. In order to suppress the masking behavior of the peaks in equation (4), log compression was later proposed [10]. This method of extraction is simpler when compared with MODGD because the group delay feature extraction algorithm does not depend on any empirical parameters and it is data independent [10].

2.2. Regularization of GD features

As indicated in equation (4), smaller values of $|S(f)|^2$ lead to large-amplitude spikes in the group delay function which can commonly create large variability in $G_s(f)$ as shown in Figure 1(a). Here, we achieve a smoother estimate, $G_{LS}(f)$, using least squares regularization. Rewriting equation (4) in a matrix-vector notation over a window of length $M$

$$n \approx G_{LS}(f)d$$  \hspace{1cm} (6)

where $n$ and $d$ are the numerator and denominator of equation (4) calculated at consecutive points along a frequency-domain window of length $M$ as in equation (9) and (10) respectively.

$$n(f_k) = X(f_k)R F\{tx(t)\}_R + X(f_k)\gamma F\{tx(t)\}_I$$  \hspace{1cm} (7)

$$d(f_k) = |S(f_k)|^2$$  \hspace{1cm} (8)

$$n = \begin{bmatrix} n(f_k) \\ \vdots \\ n(f_k-M) \end{bmatrix}$$  \hspace{1cm} (9)

$$d = \begin{bmatrix} |S(f_k)|^2 \\ \vdots \\ |S(f_k-M)|^2 \end{bmatrix}$$  \hspace{1cm} (10)

Then

$$G_{LS}(f) = \left[d^T d\right]^{-1}d^T n$$  \hspace{1cm} (11)

where the subscript $k$ denotes the frequency index. As shown in Figure 1(c), following least squares regularization, the dynamic range of variation is reduced. Compared with log compression (Fig. 1(b)), $G_{LS}$ preserves the relative significance of the GD spectral peaks more effectively. Significant additional compression can be attained by incorporating log compression, without reducing the significance of the peaks relative to the remainder of the GD spectrum as shown in Figure 1(d).

2.3. Least square group delay (LSGD) feature extraction

LSGD features can thus be computed using equation (11) for every 20 ms frames of speech. In order to decorrelate the resulting GD features, Discrete Cosine Transform (DCT) is applied to the GD function and the first 14 DCT coefficients are taken as feature [7]. The linear decorrelation provided by the DCT, allows the use of diagonal covariances in modelling the speaker vector distribution [12]. If log compression is desired, the absolute value is taken before the compression. This is a
3. Feature-domain speaker dependency experiments

As a preliminary investigation of the degree to which speakers can be separated using group delay features, the invariant cluster separation index \( J \) was employed as an evaluation metric:

\[
J = \text{trace} (S_w^{-1}S_B)
\]

where \( S_w \) and \( S_B \) are the within-cluster and between-cluster scatter matrix for a data matrices whose rows \( X_i \), \( i = 1, 2, \ldots, N \) comprise features of dimensionality \( d \) and overall mean \( m \) respectively. The data are partitioned into \( N_c \) clusters representing \( N_c \) speakers, each with \( N_j \) features and mean \( m_j \) for the \( j \)th speaker.

\[
S_w = \sum_{j=1}^{N_c} \sum_{i=1}^{N_j} (X_i - m_j)^T (X_i - m_j) \\
S_B = \sum_{j=1}^{N_c} (N_j (m_j - m)^T (m_j - m))
\]

Experiments were conducted on the GD features from various extraction techniques on NIST 2001 SRE. The \( J \)-scores for various GD extraction techniques are shown in Figure 2. These results suggest that a window length of \( M = 3 \) provides a good trade-off between regularizing the GD estimate and preserving the frequency resolution. A longer window gives a smoother and more robust estimate but blurs the frequency resolution and introduces unwanted correlations between samples as shown in Figure 1(d).

4. Speaker recognition experiments

4.1. Comparison of group delay features for NIST2001 SRE

Speaker recognition experiments were conducted using the NIST 2001 database and a subset of the NIST 2008 SRE database (Telephone/Telephone). The NIST 2001 development database consists of 38 male speakers and 22 female speakers. The evaluation database comprises 74 male speakers and 100 female speakers for training, 850 male speakers and 1188 female speakers for testing. The training time for each speaker was 2 minutes and the testing segment duration was less than 60 seconds.

Table 1: Comparison of GD feature extraction techniques for speaker recognition on the NIST 2001 SRE database

<table>
<thead>
<tr>
<th>Features</th>
<th>Window Size (M)</th>
<th>EER (%)</th>
<th>Fused EER (%) with MFCCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>-</td>
<td>7.64</td>
<td>-</td>
</tr>
<tr>
<td>MODGD [12]</td>
<td>-</td>
<td>13.35</td>
<td>7.70</td>
</tr>
<tr>
<td>LogGD [10]</td>
<td>-</td>
<td>11.73</td>
<td>7.54</td>
</tr>
<tr>
<td>LSGD</td>
<td>3</td>
<td>17.86</td>
<td>7.45</td>
</tr>
<tr>
<td>LogLSGD</td>
<td>3</td>
<td>10.01</td>
<td>7.16</td>
</tr>
<tr>
<td>LogLSGD</td>
<td>5</td>
<td>10.16</td>
<td>7.26</td>
</tr>
<tr>
<td>LogLSGD</td>
<td>7</td>
<td>10.26</td>
<td>7.21</td>
</tr>
</tbody>
</table>

The back-end of the recognition system for the NIST 2001 database was based on statistical modeling using Gaussian Mixture Models (GMMs). Initially, two gender-dependent UBMs were created with 512 GMMs. For UBM creation, the development set was used. Then the training data from the evaluation set was used to adapt speaker models from the UBM. Finally, the system was tested with the testing data of the evaluation set, by detecting the target speaker as the model having the maximum likelihood for the given test segment.

NIST evaluations use the detection cost function (DCF) and equal error rate (EER) as their primary performance measure. The DCF is defined as a weighted sum of the miss and false alarm probabilities. A DET (detection error trade off) curve can also be produced using the decision scores of the experiment.

Results for speaker recognition experiments based on the NIST 2001 SRE database for MFCC, MODGD, log compressed group delay (LogGD) [10], LSGD and log compressed least square group delay (LogLSGD) are given in Table 1 and Figure 3. The LogLSGD feature with \( M = 3 \) gave the best performance among the different approaches. This is mainly because by employing least square approach prior to log compression, we could achieve a smoother estimate as compared with log compression alone, while preserving the small variation as shown in Figure 1(d). However there is trade-off between a short and long window length. As was also found in Section 3, the performance degrades slightly for longer windows. Initial experiments with weighted LS did not provide any further...
improvement over LS.

4.2. LSGD performance for NIST2008 Tel/Tel

The back-end of the recognition system for the NIST 2008 Tel/Tel database was based on GMM-SVM. This system used GMM supervectors to construct kernels of support vector machines (SVMs). Given a speaker’s speech data, a GMM is estimated by using MAP adaptation on the means of the UBM. The means of mixture components in the GMM are then concatenated to form a GMM supervector, which is used as an SVM kernel. Variability within the evaluation data arises from various sources, notably from recordings using different handsets for the telephone data, which should be removed to improve system performance. Nuisance Attribute Projection (NAP) was performed to project out the nuisance subspace from the model space so that the remaining space can better characterize the speaker [14].

For the NIST 2008 Tel/Tel evaluation, we used the NIST 2004 SRE data as the background training data, for training the universal background models (UBMs) as well as composing the set of background speakers in SVM training. A total of 3499 male speakers and 4644 female speakers from NIST 2004 SRE data were used to derive the Nuisance Attribute Projection (NAP) matrix. For the T-Norm, 75 speakers from NIST 2005 SRE data were selected for each gender. Results from this experiment showed that the improvements discussed in Section 4.1 were also found for the more contemporary NIST2008 database, where LogLSGD improved on a 11.65% EER MFCC baseline to 9.45% after fusion as shown in Figure 4. When the LogLSGD and MFCC were further fused with FM features (extracted as described in [15]), the EER dropped to 8.28% (MFCC fused with FM produced an EER of 9.11%). These results provide strong encouragement for the pursuit of alternative features to complement MFCCs.

5. Conclusion

In this paper, we have proposed an alternative complementary feature extraction method to reduce the variability of GD features derived from the speech spectrum. Interestingly, the proposed LogLSGD feature successfully reduces the dynamic range and also appears to retain the fine structure of the group delay. Least squares regularization is a simple and effective way to reduce the dynamic range of MODGDF features, by alleviating the ill-conditioning of the MODGDF calculation due to strong excitation components. Other regularisation techniques might be expected to give similar results, but this is left for future study. Evaluation on the NIST 2008 telephone database using a fusion of LogLSGD-based and with MFCC-based subsystems, demonstrated relative improvements of 18% over the performance of an MFCC-only system. This strongly supports the hypothesis that these two subsystems carry complementary information.

6. References