Correlation-Based Similarity Between Signals for Speaker Verification with Limited Amount of Speech Data

Dhananjaya N. and B. Yegnanarayana

Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai 600 036, India
{dhanu, yegna}@cs.iitm.ernet.in

Abstract. In this paper, we present a method for speaker verification with limited amount (2 to 3 secs) of speech data. With the constraint of limited data, the use of traditional vocal tract features in conjunction with statistical models becomes difficult. An estimate of the glottal flow derivative signal which represents the excitation source information is used for comparing two signals. Speaker verification is performed by computing normalized correlation coefficient values between signal patterns chosen around high SNR regions (corresponding to the instants of significant excitation), without having to extract any further parameters. The high SNR regions are detected by locating peaks in the Hilbert envelope of the LP residual signal. Speaker verification studies are conducted on clean microphone speech (TIMIT) as well as noisy telephone speech (NTIMIT), to illustrate the effectiveness of the proposed method.

1 Introduction

The amount of speech data available for automatic recognition of speakers by a machine is an important issue that needs attention. It is generally agreed upon that human beings do not require more than a few seconds of data to identify a speaker. Popular techniques giving the best possible results require minutes of data, and higher the amount of data higher is the performance. But the speaker verification performance is seen to reduce drastically when the amount of data available is only a few seconds of the speech signal. This has to do with the features chosen and the modeling techniques employed. Mel-frequency cepstral coefficients (MFCCs), the widely used features, characterize the shape and size of the vocal tract of a speaker and hence are representatives of both the speaker as well as the sound under consideration. Considering the fact that the vocal tract shapes are significantly different for different sounds, the MFCCs vary considerably across sounds within a speaker. Apart from using vocal tract features, the popular techniques for speaker verification employ statistical methods for modeling a speaker. The performance of these statistical techniques is good as long as there are enough examples for the statistics to be collected. In this direction,

B. Gunsel et al. (Eds.): MRCS 2006, LNCS 4105, pp. 17–25, 2006.

exploring the feasibility of using excitation source features for speaker verification gains significance. Apart from adding significant complementary evidence to the vocal tract features, the excitation features can act as a primary evidence when the amount of speech data available is limited.

The results from NIST-2004 speaker recognition evaluation workshop [1] show that a performance of around 12% EER (equal error rate) is obtained when a few *minutes* of speech data is available. These techniques typically use the vocal tract system features (MFCCs or mel frequency cepstral coefficients) and statistical models (GMMs or Gaussian mixture models) for characterizing a speaker. Incorporation of suprasegmental (prosodic) features computed from about an hour of data, improves the performance to around 7% EER [1]. At the same time, the performance reduces to an EER of around 30%, when only *ten* seconds of data is available. One important thing to be noted is that the switchboard corpus used is large, and has significant variability in terms of handset, channel and noise.

In forensic applications, the amount of data available for a speaker can be as small as a few phrases or utterances, typically recorded over a casual conversation. In such cases, it is useful to have reliable techniques to match any two given utterances. The availability of only a limited amount of speech data, makes it difficult to use suprasegmental (prosodic) features, which represent the behavioral characteristics of a speaker. Also, the use of statistical models like Gaussian mixture models (GMMs) along with the popular mel frequency cepstral coefficients (MFCCs) becomes difficult, owing to nonavailability of enough repetitions of different sound units reflecting different shapes of the vocal tract. These constraints force one to look into anatomical and physiological features of the speech production apparatus that do not vary considerably over the different sounds uttered by a speaker. Some of the available options include the rate of vibration of the vocal folds (F_0, the pitch frequency), the length of the vocal tract (related to the first formant F_1), and parameters modeling the excitation source system [2,3]. Some of the speaker verification studies using the excitation source features are reported in [3] [4] [5]. Autoassociative neural network (AANN) models have been used to capture the higher order correlations present in the LP residual signal [4] and in the residual phase signal (sine component of the analytic signal obtained using the LP residual) [5]. These studies show that reasonable speaker verification performance can be achieved using around *five* seconds of voiced speech. Speaker verification studies using different representations of the glottal flow derivative signal are reported in [3]. Gaussian mixture models are used to model the speakers using around 20 to 30 seconds of training data. The use of MFCCs computed from the GFD signal gives a good performance (95% correct classification) for clean speech (TIMIT corpus), as compared to (around 70%) using parameters modeling the coarse and fine structures of the GFD signal. The performance is poor (25% using MFCCs) for the noisy telephone speech data (NTIMIT).

In this paper, we outline a method for speaker verification that compare two signals (estimates of the GFD signals), without having to extract any further

parameters. In Section 2, a brief description of estimating the glottal flow derivative signal is given. Section 3 describes a correlation-based similarity measure for comparing two GFD signals. Some of the issues in the speaker verification experiments are discussed in Section 4. The performance of the speaker verification studies are given in Section 5, followed by a summary and conclusions in the final section 6.

2 Estimation of the Glottal Flow Derivative Signal

The speech production mechanism in human beings can be approximated by a simple cascade of an excitation source model, a vocal tract model and a lip radiation model [2]. The vocal tract model can be approximated by an all-pole linear filter using linear prediction (LP) analysis, and the coupling impedance at the lip is characterized by a differentiator. A reasonable estimate of the glottal flow derivative (GFD) signal can be obtained by using a two-stage filtering approach. First, the speech signal is filtered using the LP inverse filter to obtain the LP residual signal. The LP residual signal is then passed through an integrator to obtain an estimate of the GFD signal.

Fig. 1 shows the estimated GFD signals for five different vowels /*a*/, /*i*/, /*u*/, /*e*/ and /*o*/ for two different male speakers. The signals have been aligned

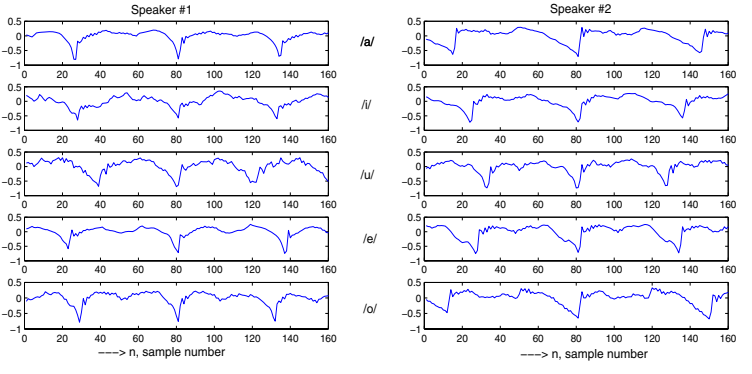

Fig. 1. Estimates of the glottal flow derivative signal for five different vowels of two different speakers

at sample number 80, corresponding to an instant of glottal closure (GC). A close observation of the signals around the instants of glottal closure shows that there exists a similar pattern among the different sounds of a speaker. The objective is to cash on the similarity between signal patterns within the GFD signal of a speaker, while at the same time bring out the subtle differences across speakers. The normalized correlation values between signal patterns around the high SNR regions of the glottal flow derivative signal are used to compare two GFD signals. Approximate locations of the high SNR glottal closure regions

(instants of significant excitation) are obtained by locating the peaks in the Hilbert envelope of the LP residual signal, using the average group delay or phase-slope method outlined in [6].

3 Correlation-Based Similarity Between Two GFD Signals

The similarity between any two signal patterns $r_1[n]$ and $r_2[n]$ of equal lengths, say N samples, can be measured in terms of the cross-correlation coefficient

$$\rho(r_1[n], r_2[n]) = \frac{\sum_{n=0}^{N-1} (r_1[n] - \mu_1)(r_2[n] - \mu_2)}{(\sum_{n=0}^{N-1} (r_1[n] - \mu_1)^2)^{1/2} (\sum_{n=0}^{N-1} (r_2[n] - \mu_2)^2)^{1/2}} \tag{1}$$

where μ_1 and μ_2 are the mean values of $r_1[n]$ and $r_2[n]$. The values of the cross-correlation coefficient ρ lie in the range [-1 to +1]. A value of $\rho = +1$ indicates a perfect match, and $\rho = -1$ indicates a 180^o phase reversal of the signal patterns. Any value of $|\rho| \rightarrow 0$ indicates a poor match. While operating on natural signals like speech, the sign of the cross-correlation coefficient is ignored, as there is a possibility of a 180^o phase reversal of the signal due to variations in the recording devices and/or settings.

Let $x[n]$ and $y[n]$ be any two GFD signals of lengths N_x and N_y, respectively, which need to be compared. Let $\mathcal{T}_x = \{\tau_0, \tau_1, \ldots, \tau_{N_1-1}\}$ and $\mathcal{T}_y = \{\tau_0, \tau_1, \ldots, \tau_{N_2-1}\}$ be the approximate locations of the instants of glottal closure in $x[n]$ and $y[n]$, respectively. Let $z[n] = x[n] + y[n - N_x]$ be a signal of length $N_z = N_x + N_y$ obtained by concatenating the two signals $x[n]$ and $y[n]$, and $\mathcal{T}_z = \{\mathcal{T}_x, \mathcal{T}_y\} = \{\tau_0, \tau_1, \ldots, \tau_{N-1}\}$ be the concatenated set of locations of the reference patterns, where $N = N_1 + N_2$. Let $\mathcal{R} = \{r_0[n], r_1[n], \ldots, r_{N-1}[n]\}$ be the set of signal patterns of length N_r chosen symmetrically around the corresponding GC instants in $\mathcal{T}_z$. Now, for each reference pattern $r_i[n] \in \mathcal{R}$, the similarity values with all other patterns in $\mathcal{R}$ is computed, to give a sequence of $\cos\theta$ values

$$c_i[j] = \max_{-N_\tau \leq k \leq +N_\tau} |\rho(r_i[n], z[n - \tau_j + k])| \qquad j = 0, 1, \ldots, N-1 \tag{2}$$

$$C = \{c_i[n]\} \quad i = 0, 1, \ldots, N-1 \tag{3}$$

where N_τ represents the search space around the approximate locations specified in $\mathcal{T}_z$.

The first (N_1) $\cos\theta$ plots (or rows) in C belong to patterns from $x[n]$, and hence are expected to have a similar trend (relative similarities). They are combined to obtain an average $\cos\theta$ plot $\bar{c}_x[n]$. Similarly, the next ($N_2 = N - N_1$) $\cos\theta$ plots are combined to obtain $\bar{c}_y[n]$. Figs. 2(a) and 2(b) show typical plots of $\bar{c}_x[n]$ and $\bar{c}_y[n]$ for a genuine and an impostor test, respectively. It can be seen that $\bar{c}_x[n]$ and $\bar{c}_y[n]$ have a similar trend when the two utterances are from

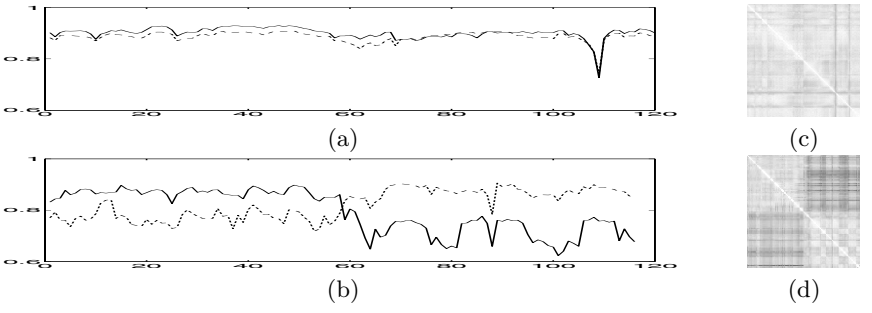

Fig. 2. Average $\cos\theta$ plots $\bar{c}_x[n]$ (solid line) and $\bar{c}_y[n]$ (dashed line) for a typical genuine and impostor test ((a) and (b)). Intensity maps of the similarity matrices for a typical genuine and impostor test ((c) and (d)).

the same speaker, and have an opposite trend when the speakers are different. The similarity matrix C may also be visualized as a 2-D intensity map. Typical similarity maps for an impostor (different speakers) test and a genuine (same speaker) test are shown in Figs. 2(c) and 2(d). The 2-D similarity matrix can be divided into four smaller blocks as

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} \tag{4}$$

where C_{xx} and C_{yy} are the similarity values among patterns within the train and test utterances, respectively, and C_{xy} and C_{yx} are the similarity values between patterns of the train and test utterances. The similarity values in C_{xx} and C_{yy} are expected to be large (more white), as they belong to patterns from the same utterance. The values in C_{xy} and C_{yx}, as compared to C_{xx} and C_{yy} are expected to be relatively low (less white) for an impostor, and of similar range for a genuine utterance. As can be seen from Fig. 2, the $\cos\theta$ values lie within a small range (around 0.7 to 0.9), and hence the visual evidence available from the intensity map is weak. Better discriminability can be achieved by computing a second-level of similarity plots $S = \{s_i[n]\}$, $i = 0, 1, \ldots, N-1$, where $s_i[j] = \rho(c_i[n], c_j[n])$, $j = 0, 1, \ldots, N-1$. The second-level average $\cos\theta$ plots $\bar{s}_x[n]$ and $\bar{s}_y[n]$ and the second-level similarity map are shown in Fig. 3. A final similarity measure between the two signals $x[n]$ and $y[n]$ is obtained as

$$s_f = \rho(\bar{s}_x[n], \bar{s}_y[n]) \tag{5}$$

Now, if both the signals $x[n]$ and $y[n]$ have originated from the same source (or speaker), then $\bar{s}_x[n]$ and $\bar{s}_y[n]$ have similar trend, and $s_f \to +1$. In ideal cases, $s_f = +1$, when $x[n] = y[n]$, for all n. On the other hand, if $x[n]$ and $y[n]$ have originated from two different sources, then $\bar{s}_x[n]$ and $\bar{s}_y[n]$ have opposite trends and $s_f \to -1$. In ideal cases, $s_f = -1$, when $x[n] = -y[n]$, for all n.

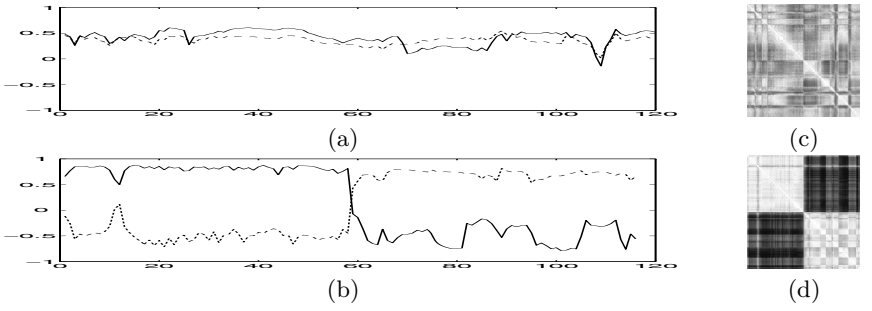

Fig. 3. Second-level average $\cos\theta$ plots $\bar{s}_x[n]$ (solid line) and $\bar{s}_y[n]$ (dashed line) plots for a typical genuine and impostor test ((a) and (b)). Intensity maps of the second-level similarity matrices for a typical genuine and impostor test ((c) and (d)).

4 Speaker Verification Experiments

The speaker verification task involves computation of a similarity measure between a train utterance (representing a speaker identity) and a test utterance (claimant), based on which a claim can be accepted or rejected. Estimates of the GFD signals for both the train and test utterances, say $x[n]$ and $y[n]$, are derived as described in Section 2. The correlation-based similarity measure s_f given by Eqn. (5) is computed as outlined in Section 3. A good match gives a positive value for s_f tending toward $+1$, while a worst match (or a best impostor) gives a negative value tending toward -1.

The width of the reference frame T_r ($T_r = N_r/F_s$ where F_s is the sampling rate) is a parameter which can affect the performance of the verification task. A reasonable range for T_r is between 5 ms to 15 ms, so as to enclose only one glottal closure region. In our experiments, a value of T_r =10 ms is used. The signal patterns are chosen around the instants of glottal closures, and errors in the detection of the instants of glottal closures (e.g. secondary excitations and

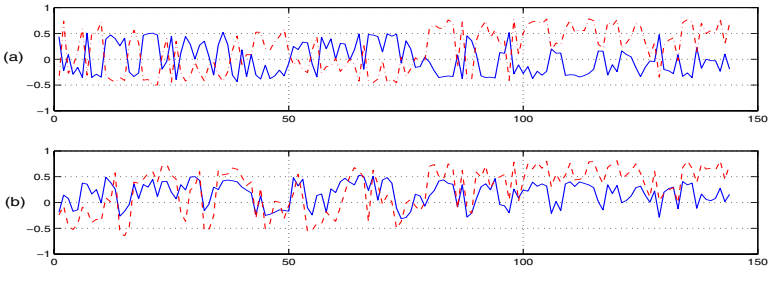

Fig. 4. Consolidated similarity plots $\bar{s}_x[n]$ (solid line) and $\bar{s}_y[n]$ (dashed line) for (a) an impostor and (b) a genuine claim

unvoiced regions) result in spurious patterns. Such spurious patterns are eliminated by computing the second-level similarity matrices S_x and S_y separately for $x[n]$ and $y[n]$, and picking a majority of patterns which have similar trends. A few spurious patterns left out do not affect the final similarity score (genuine or impostor). The advantage of using the relative similarity values $\bar{s}_x[n]$ and $\bar{s}_y[n]$ for computing the final similarity measure s_f, can be seen from the plots in Fig. 4. The relative similarities have an inverted trend for an impostor, while the trend is similar for a genuine claim.

5 Performance of Speaker Verification Studies

The performance of the signal matching technique for speaker verification was tested on clean microphone speech (TIMIT database), as well as noisy telephone

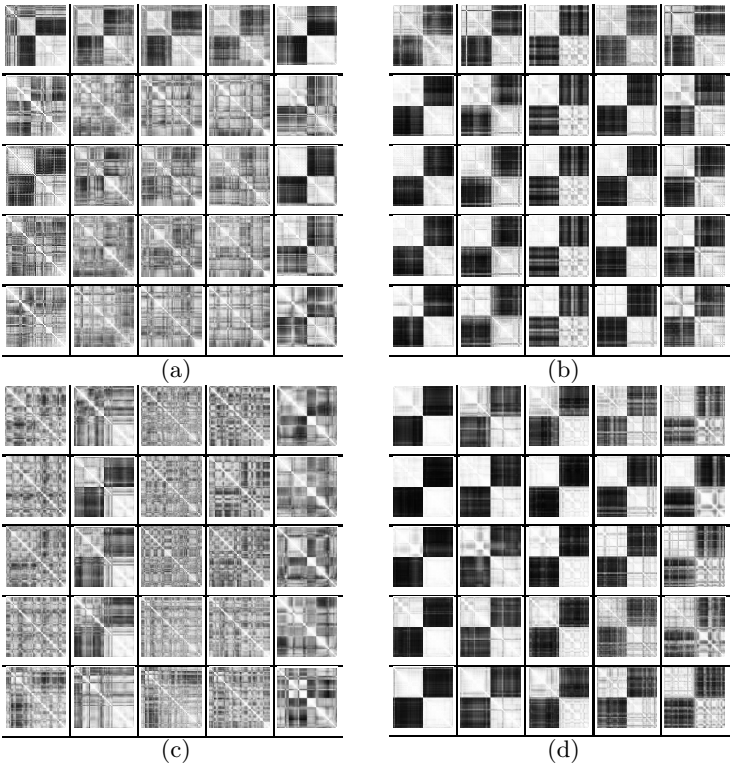

Fig. 5. (a) Intensity (or similarity) maps for twenty five genuine tests. Five different utterances of a speaker (say S_1) are matched with five other utterances of the same speaker. (b) Intensity maps for twenty five impostor tests. Five different utterances of speaker S_1 matched against five different utterances (five columns of each row) of five different speakers. (c) and (d) Genuine and impostor tests for speaker S_2, similar to (a) and (b).

speech data (NTIMIT database). The datasets in both cases consisted of twenty speakers with ten utterances (around 2 to 3 secs) for each, giving rise to a total of 900 genuine tests and 18000 impostor tests. Equal error rates (EERs) of 19% and 38% are obtained for the TIMIT and NTIMIT datasets, respectively. Several examples of the intensity maps for genuine and impostor cases are shown in Fig. 5. It can be seen from Fig. 5(a) that the first train utterance (first row) gives a poor match with all five test utterances of the same speaker. Similar are the cases for the fifth test utterance (fifth column) of Fig. 5(a), and the second test utterance (second column) of Fig. 5(c). Such behaviour can be attributed to poorly uttered speech signals. The performance can be improved when multiple train and test utterances are available. At the same time, it can be seen from Figs. 5(b) and (d) that there is always significant evidence for rejecting an impostor. The same set of similarity scores (i.e., scores obtained by matching one utterance at a time) was used to evaluate the performance when more number of utterances (three train and three test utterances) are used per test. All possible combinations of three utterances against three other were considered. The nine different similarity scores available for each verification are averaged to obtain a consolidated score. The EERs improve to 5% for TIMIT and 27% for NTIMIT datasets. The experiments and results presented in this paper are only to illustrate the effectiveness of the proposed method. More elaborate experiments on NIST datasets need to be conducted to compare the effectiveness of the proposed method as against other popular methods.

6 Summary and Conclusions

The emphasis in this work has been on exploring techniques to perform speaker verification when the amount of speech data available is limited (around 2 to 3 secs). A correlation-based similarity measure was proposed for comparing two glottal flow derivative signals, without needing to extract any further parameters. Reasonable performances are obtained (for both TIMIT and NTIMIT data), when only one utterance is available for training and testing. It was also shown, that the performance can be improved when multiple utterances are available for verification. While this work provides a method for verifying speakers from limited speech data, it may provide significant complementary evidence to the vocal tract based features when more data is available. The proposed similarity measure, which uses the relative similarity among patterns in the two signals, can be generalized for any sequence of feature vectors, and any first-level similarity measure (instead of $cos\theta$).

References

1. NIST-SRE-2004: One-speaker detection. In: Proc. NIST Speaker Recognition Evaluation Workshop, Toledo, Spain (2004)
2. Ananthapadmanabha, T.V., Fant, G.: Calculation of true glottal flow and its components. Speech Communication (1982) 167–184

3. Plumpe, M.D., Quatieri, T.F., Reynolds, D.A.: Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech and Audio Processing **7** (1999) 569–586
4. Yegnanarayana, B., Reddy, K.S., Kishore, S.P.: Source and system features for speaker recognition using AANN models. In: Proc. Int. Conf. Acoustics Speech and Signal Processing. Volume 1., Salt Lake city, Utah, USA (2001) 409–412
5. Murthy, K.S.R., Prasanna, S.R.M., Yegnanarayana, B.: Speaker-specific information from residual phase. In: Int. Conf. on Signal Processing and Communications, SPCOM-2004, Bangalore, India (2004)
6. Smits, R., Yegnanarayana, B.: Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech and Audio Processing **3** (1995) 325–333