Automated Speaker Recognition Using Compressed Temporal-Spectral Dynamics Information of Password Spectrograms

Amitava Das¹ and Gokul Chittaranjan²

¹ Microsoft Research Lab – India; 196/36 2nd Main; Sadashivnagar; Bangalore India 560 080.
² Student intern at MSR-India
amitavd@microsoft.com

Abstract

Prevalent speaker recognition methods use only spectral-envelope based features such as MFCC, ignoring the rich speaker identity information contained in the temporal-spectral dynamics of the entire speech signal. We propose a new feature for speaker recognition called compressed spectral dynamics (CSD) which effectively captures such spectral dynamics and the inherent speaker identity. The discriminative power of CSD allows the classification part to remain simple. The proposed method, a simple nearest neighbor classifier using CSD, delivers performance competitive to conventional MFCC+DTW based text-dependent speaker recognition methods at significantly reduced complexity.

Index Terms: Speaker Verification; Temporal & Spectral Dynamics; New Feature; Nearest Neighbor Classifier;

1. Introduction

The majority of the prevalent speaker recognition methods [1-5, 8-10] use speech spectral envelope parameters such as Mel-Frequency Cepstral Coefficient (MFCC) as the main feature for classification. MFCC offers a compact representation of the speech spectral envelope or the impact of the vocal tract shape in rendering a particular sound. MFCC is quite useful for speech recognition. But for speaker recognition it is questionable whether MFCC is the best and a complete representation of speaker identity, mainly for the following reasons. There is significant speaker identity information in the excitation part of the signal which is completely missing when only MFCC is used. Secondly, there is significant temporal dynamics in the speech signal. The traditional MFCC-plus-derivative representation captures only a highly localized portion of that dynamics. It is ironic that the same MFCC feature is considered for speaker-independent speech recognition and for the speaker recognition task as well.

There is a wealth of speaker-identity information in the temporal-spectral dynamics contained in the complete speech signal — much more than the compressed envelope-only representation offered by MFCC. There have been a number of recently proposed speaker recognition methods [6,7] looking beyond MFCC. In this paper, we propose a new approach of automated speaker recognition using spectrogram. However, it is important to look at prevalent speaker recognition methods briefly to understand why speech dynamics are important for speaker recognition.

For automated speaker recognition, the two prevalent methods are: a) text-independent (TI) and b) text-dependent(TD). Text-independent methods assume that the password the user is uttering can be anything. TI methods pay no attention to feature dynamics and treat the sequence of extracted features from the speech utterance not as a sequence of symbols but as a bag of symbols. Speaker models in TI methods are distributions in the feature space, modeled by VQ codebooks [9-11] or by Gaussian mixture models [1, 14] built from training data. Such distributions are often overlapping. During testing, the task of TI speaker recognition amounts to finding which speaker model (distribution) the test feature-vector-set came from. The way a person speaks a certain phrase, captures a lot of his/her speaking style (i.e. the identity), in the co-articulation of various sound units. TI methods totally miss this important aspect of speaker identity. As a result, the performance accuracies of TI methods are much lower than those of the TD methods.

Text-dependent (TD) speaker recognition methods [2-5, 8] on the other hand exploit the feature dynamics to capture the identity of the speaker. TD methods assume the utterance of a certain password by the speaker and compare the feature vector sequence of the test utterance with the “feature-dynamics-model” of all the speakers. Such feature-dynamics based speaker models can simply be the stored templates of feature vector sequence, collected during training, or they can be HMMs trained by a large number of passwords uttered by the speaker. TD methods therefore use dynamic classification methods such as Dynamic Time Warping (DTW) [8,13] or HMM decoders [3,4] to match the input feature sequence with the HMM-based speaker model. Table 1 below summarizes the key aspects of TI and TD methods in a qualitative manner.

We propose a new TD speaker recognition method which utilizes the entire speech signal. The novelties of the proposed method are: a) it exploits the temporal-spectral dynamics contained in the speech signal and looks beyond the spectral envelope information as in conventional methods, b) it captures the above temporal-spectral dynamics, indicative of a speaker’s style of speaking, in a compact signature vector called Compressed Spectral Dynamics or CSD vector, and c) it proposes a simple nearest-neighbor classifier-based TD speaker-recognition method which performs as good as conventional TD methods but at a significantly reduced complexity.

We compared our method with a conventional MFCC+DTW based TD method as in [8] as baseline and the results show the proposed method is quite competitive with respect to the baseline method, in terms of performance and complexity. The paper is organized as follows: Section 2 present the proposed method. Section 3 presents the database and experimental evaluation details and results finally section 4 presents the conclusions and future directions.
2. Proposed CSD Feature for Speaker Recognition

We claim that the temporal dynamics in speech signal, which often reveals speaker identity, can be efficiently captured in a properly resolved spectrogram. A spectrogram is a 2-D representation of the 1-D speech signal in which absolute magnitude of the spectra of successive speech segments are laid out to create a 2-D Frequency-Time (x-axis) representation. If a speech utterance has M number of segments of length T samples (typically T = 160-256) and each segment is converted to an N-point spectral magnitude vector, then a N x M size spectrogram is created. Spectrograms have been a highly useful and popular means of studying and analyzing the acoustic phonetic information in speech signal. There are a number of experts who are trained to read spectrograms to identify the content and the speaker. Today spectrograms are widely used for forensic and legal use of speaker recognition. These are however all manual methods. We propose an automated method which can efficiently capture the temporal-spectral dynamics in spectrogram and utilize it for speaker recognition overcoming the two well-known barriers described next.

![Figure 1. Examples showing spectrograms of the same password uttered by the target speaker (two versions) and an imposter](image)

A NxM size spectrogram is quite "unfriendly" for automated classification. This is because of the huge amount of data (NxM numbers) to process and store. For example, a system using spectrograms with N=512 M=100 (on average) and four password templates per speaker will need to store 202400 numbers per speaker. A more significant problem is the variability of M with the duration of the speech segment, which makes the comparison of two spectrograms of size NxM1 and NxM2 a difficult task. If we use a DTW type method to handle this, it will be incredibly complex. These two problems are solved by the proposed Compressed Spectral Dynamics feature extraction method.

2.1. The Compressed Spectral Dynamics Feature

We propose a new feature for speaker recognition derived from spectrogram called "compressed spectral dynamics" or CSD, which is essentially a fixed-dimension vector derived from the speech signal using the following steps:

**Speech password** → **Spectrogram** SG → **Apply Transform**, W: T = \( W(S) \) → CSD = \( f_{select}(T) \), where \( W \) is a suitable transform and \( f_{select} \) is a way of selecting \( K \) elements from \( T \) (see Figure 2)

![Figure 2. CSD extraction using DCT](image)

To compare two passwords and hence two spectrograms, \( SG_a(NxM1) \) and \( SG_b(NxM2) \), we follow these steps: a) resizing \( SG_a \) to \( SG_b \) to the size of \( SG_b \) using 2-D image interpolation methods, b) Extract CSD from \( SG_a' \) (\( NxM1 \)) and \( SG_b \), and c) Find distance between the two CSD’s as shown below.

\[
\text{dist}(\text{utterance-A, utterance-B}) = \text{dist}(SG_a, SG_b) \\
\quad = \text{dist}(A, B') = \text{dist}(CSD_a, CSD_b) \quad \ldots (1)
\]

A suitable choice of transform ensures that the crucial distance property (Eqn. 2 shown below) is maintained:

\[
\text{dist}(A, B) < \text{dist}(A, C) \Rightarrow \text{dist}(CSD_a, CSD_b) < \text{dist}(CSD_a, CSD_c) \quad \ldots (2)
\]

We have chosen Discrete Cosine Transform [12] as the transform which preserves the above distance property as well as packs most of the information in a small set of transform coefficients allowing us to define a "f.select" function as illustrated in Figure 2 below. We omit the DC value and keep the top \( K=m^2-1 \) coefficients (Figure 2) in a zigzag scan to create the K-dimension CSD vector.
Examples of Spectrograms of passwords of a target speaker and an imposter are shown in Figure 1. The CSDs of the same passwords are shown in Figure 3. Note that now the entire speech utterance is represented by a fixed K-dimension CSD vector.

![Figure 3: CSD's of the same spectrograms shown in Figure 1](image)

Figure 3 illustrates a few important things. First of all, the “variable X dimension problem” of spectrograms is gone as they are converted into fixed dimension CSD vectors. Also note that the spectrograms 1 and 3 are similar looking even though their x-dimensions are different and spectrogram 1 and 3 are different-looking. Such discrimination is preserved in the CSD domain -- d(CSD1, CSD2) is less than d(CSD1, CSD2). Thus spectrograms and CSDs can capture speaker identity well by preserving the speaker-specific temporal-spectral dynamics in the speech signal. Also note that a conventional MFCC+DTW method need to store ~3900 numbers per password. In contrast, the proposed CSD-based method needs to store only K(typically K=35) numbers per password. This also reduces run-time complexity compared to conventional TD-DTW methods. The proposed CSD-MSRI speaker recognition method is presented next.

### 2.2 Two-Stage Combined Static-Dynamic CSD-MSRI method for Speaker Recognition

Our proposed CSD-MSRI method has an MFCC+VQ based static classification 1st stage which picks Nbest candidates closest to the test utterance. This is followed by a CSD+nearest-neighbor based 2nd stage which looks at the speech dynamics of these Nbest candidates using CSD and makes the final decision. The key algorithms of the two stages are given below.

#### First Stage Algorithm:

**Design S CxS size VQ codebooks** \{CB1, CB2, … \}one for each of the S enrolled speakers, using D dimension MFCC as feature. Each codebook has C code vectors. Given a test utterance of M frames, we extract a D-dimension MFCC vector sequence \{F1, F2, Fm, … FM\} and follow these steps:

**Step 1:** For each speaker \( P \), compute an accumulated distance \( A(m) \) as follows:

\[ A(m) = D_1(1) + D_1(2) + \ldots + D_1(m) \]

where \( D_1(m) \) is the minimum distance of \( F_m \) to \( CB_i \), (codebook of person \( P \)); \( D_1 = \min \{ D_{i,j} \} \) where \( D_{i,j} = \| F_{m} - C_{i,j} \| \) \( j=1,2, \ldots L \), \( C_{i,j} \) being the code vector of the codebook \( CB_i \).

**Step 2:** Pick the Nbest candidates, for which \( A(M) \) are lowest, i.e. for \( N_{best}=2, \) pick the 3 speakers having the 3 lowest \( A(M) \) scores.

After 1st stage we have a selected set of candidates

\[ CP = [CP_{i} =1, 2,3, \ldots N_{best}] \]

#### Second Stage Algorithm:

For each of the candidate speaker \( CP_i \) in the \( N_{best} \) speaker set, we calculate a distance \( D_2 \) as follows:

Let \( D_2 = \text{dist}(\text{SG}_{\text{test}}, \text{SG}_{\text{template}(i,j)}) = D(CSD_{\text{test}}, CSD_{i,j}) \)

where \( SG_{\text{test}} \) and \( CSD_{\text{test}} \) are the spectrogram and CSD of the test utterance and \( \text{SG}_{\text{template}(i,j)} \) and \( CSD_{i,j} \) are the spectrogram and CSD of the \( j \)-th stored password template of the candidate speaker \( CP_i \).

Let \( D_k \) be the minimum distance over all \( D_k \) over \( j=1,2,3 \ldots N_{best} \) we create a distance array \( D = [D_k, k=1,2,3, \ldots N_{best}] \).

For Speaker Identification: Out of \( k=1,2, \ldots N_{best} \) speakers, pick the speaker for which \( D_k \) is the minimum

#### For Speaker Verification:

1. Create a modified candidate list \( CP^* = [\text{target-speaker}, N_{best} \text{speakers excluding target-speaker}], i.e. \text{the target speaker is 1st in the list.} \)
2. Calculate two distance-sets (D= [D1] as above): a) \( D_{\text{VQ}} \) using CSD as feature, and b) \( D_{\text{VQ}} \) using MFCC-VQ-distance \( A(M) \)
3. Calculate two likelihood ratios \( R_{\text{VQ}} & R_{\text{VQ}} \) as follows R= d1/d2, where d2=D(1); d1= minimum(D'), where D'=|[D2] \ldots D(N_{best})|
4. Calculate a product fusion score \( R_p = R_{\text{VQ}} \times R_{\text{VQ}} \)
5. Accept if \( R_p > TH \) else reject, where \( TH \) is some pre-computed threshold

Thus the proposed method exploits both static and dynamic classification methods and exploits envelope as well as entire speech signal information. The system parameters are: C=size of 1st stage VQ codebook, D=dimension of 1st stage VQ codebooks, Nbest=K=Dimension of CSD and P=number of password templates per speaker. Note that it is also important to choose the proper window-size and hop-size parameters for creating the appropriate type of spectrogram. We used narrowband spectrogram as it reveals harmonic structure containing speaker identity and found it to deliver higher performance than wideband ones.

### 3. Database, experimental setup, baseline system and results

The CSD-MSRI method is compared with a MFCC+DTW speaker recognition system (similar to [8]) which uses 39-dim MFCC and conventional DTW[12] with simplest local paths and r=8 as global constraint. We have not tried any optimization techniques for the baseline DTW. 4 password templates were used for each speaker for training.

For our evaluation, we needed a database with a large number of speakers in which each speaker is using unique password and each speaker is also trying to impose as a target speaker by: a) saying random password (unknown-password impostor trial) , b) saying the password of the target speaker (known-password impostor trial). We are not aware of any publicly available database which meets the above requirements. The closest one we found is LDC-YOHO, but it does not provide several versions (same and multiple sessions) of unique password per speaker. Thus, we had to create our own MSRI Speaker Recognition database. To our knowledge, this is the most comprehensive database for text-dependent speaker recognition tasks. This MSRI database is available for anyone for research usage.

The MSRI database has 370 speakers, recorded in an office environment over a period of 9 months. 20 of these users were recorded in multiple (6) sessions, separated by 4 weeks.
Each user selected a unique 4-word password. No specific effort was made to control the environmental noise. The database therefore has realistic office background conditions with SNRs ranging from ~8-45 dB. Each person made a recording of 12 to 20 versions of his/her own password as well as passwords of other users, plus some random 4-digit passwords. For the evaluation reported in this paper, we used 300 speakers; used 4 passwords for training and remaining ones for testing. This created 3581 speaker-identification trials and 8173 speaker verification trials (3581 target trials, 4592 imposter trials in which 1011 trials are known-password-imposter trials). The results for the baseline DTW and the proposed CSD-MSRI system are shown in Table 2.

<table>
<thead>
<tr>
<th>Performance Comparison</th>
<th>DTW+MFCC</th>
<th>CSD-MSRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration in % EER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Known PWD</td>
<td>0.34 %</td>
<td>0.39 %</td>
</tr>
<tr>
<td>Unknown PWD</td>
<td>0.39 %</td>
<td>0.34 %</td>
</tr>
<tr>
<td>Complexity Comparison</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage: numbers to store/speaker</td>
<td>15600</td>
<td>716</td>
</tr>
<tr>
<td>Run-time** (seconds/trial)</td>
<td>824.13</td>
<td>2.4</td>
</tr>
</tbody>
</table>

*Run-time is the average time taken to run a single test trial in a 300-speaker recognition task (Matlab running on 3.5 GHz/2GB Pentium4)

Table 2. Comparison of performance and complexity of proposed MSRI method (D=9; C=13; P=4; K=143) with baseline DTW(4 template/speaker)

4. Conclusions and Future Directions

As seen in Table 2, the proposed Compressed Spectral Dynamics feature and the two-stage CSD-MSRI method are quite effective in speaker recognition tasks. Compared to the conventional MFCC+DTW TD method, the performance of the CSD-MSRI method is competitive and the complexity is approximately 400 times lower than DTW. Note that for speaker verification trials with imposters not-knowing target passwords, MFCC+DTW method worked well (0.3392% EER), while the MSRI-CSD method produced no errors. For speaker verification trials with imposters knowing target passwords, MFCC+DTW method produced higher errors (3% EER) than the unknown-password case. Since the speech content was the same and as MFCC primarily represents speech content, the MFCC+DTW method could not separate imposters from true clients in many cases. The CSD-MSRI method worked better (0.99% EER) as it exploited the entire speech signal.

Overall, the results show the higher discrimination power of CSD over MFCC. Our experimental trials were rigorous with large variations in speakers, noise, session and content, so that the performance edge the CSD-MSRI method shows over the MFCC+DTW method do validate our claim that for speaker recognition there is merit in processing the entire speech data than using only the spectral envelope information.

To summarize, we claim that there exists significant speaker identity information in a properly resolved spectrogram. In the past, spectrograms were used manually in speaker recognition tasks, but there were no compact representation to make such information amenable to automated classification. The proposed CSD-based method enables us to overcome this hurdle and presents a feasible and efficient way to capture the entire speech dynamics.

This results in performance competitive to current text-dependent speaker recognition methods at a much lower complexity. Another novelty of the proposed method is the capturing and presentation of the complex temporal-spectral dynamics inherent in a speech segment as an “image” and eventually by a compact fixed-dimension signature vector. This opens a promising new horizon to explore this further for speech and speaker recognition and at present we are pursuing this further with newer normalization methods, transforms, and various time-feature representation methods.

5. References