Text-Dependent Speaker Recognition by Efficient Capture of Speaker Dynamics in Compressed Time-Frequency Representations of Speech

Amitava Das and Gokul Chittaranjan

Microsoft Research Lab – India; 196/36 2nd Main; Sadashivnagar; Bangalore. India 560 008
amitavd@microsoft.com
*student intern at MSR-India

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 called compressed spectral dynamics or CSD for speaker recognition based on a compressed time-frequency representations of spoken passwords which effectively captures the speaker identity. The fixed-dimension nature of the CSD allows classification to remain simple while keeping the discriminatory power of the 2D intermediate time-frequency representations. The proposed MSRI-CSD text-dependent speaker recognition method uses a simple nearest neighbor classifier and delivers performance competitive to conventional MFCC+DTW based speaker recognition methods at significantly lower complexity.

Index Terms: text-dependent speaker recognition; text-conditioning; new features; nearest neighbor classifier;

1. Introduction
Spectral features like Mel-Frequency Cepstral Coefficient (MFCC) are used to capture speaker identity in almost all of the prevalent speaker recognition methods [1-6], except a few recent approaches [7,8]. 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 feature or whether it offers a complete representation of speaker identity for the following reasons. A spectral-envelope only feature like MFCC completely misses the significant speaker identity information present in the excitation part of the spectral signal. Also, if we consider speaker-identity being mostly expressed in the temporal-spectral dynamics of the spoken password, 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.

In this work, we propose several novel approaches to capture speaker-identity for text-dependent speaker recognition. The identity or the speaking style of a person is mostly expressed in the speech dynamics primarily exhibited in the co-articulation of various sound units. We present a novel way to capture such speaker-specific dynamics, exhibited in various time-frequency representations of their password, by a compact feature vector defined as “Compact Spectral Dynamics” or CSD. CSD representation allows one to capture such speaker specific dynamics in a single fixed-dimension feature vector. In this paper, we illustrate the use of CSD in the following two time-frequency representations of speech: a) the conventional spectrogram and b) a new time-frequency speaker-identity representation called Sinogram, based on the sinusoidal representation of speech [12]. Both of these two time-frequency representations offer significantly richer information about speaker identity than MFCC. Such time-frequency representations are then compressed to form the final fixed-dimension vector representation of the spoken password which we call CSD.

Conventional text-dependent (TD) speaker recognition methods extract the MFCC sequence from the spoken password. This MFCC sequence will have different length from utterance to utterance and from speaker to speaker. As a result, the conventional TD methods need to apply a dynamic classification methods such as HMM [3,4] or DTW[6,11] to handle the variable-dimensional MFCC sequence. The proposed CSD approach offers a distinct advantage over conventional TD methods. Since the CSD vector is a fixed-dimensional, we can employ a simple nearest-neighbor classifier. This leads to a significant reduction in complexity and storage requirements while delivering performance competitive to conventional methods. These advantages of the proposed CSD approach is illustrated by a simple nearest-neighbor classifier based MSRI-CSD text-dependent speaker recognition method presented in this paper.

Our paper is organized as follows: Section 2 presents the proposed compressed time-frequency representation of spoken passwords. Section 3 presents the details of MSRI-CSD speaker recognition method. Section 4 presents the details of experimental trials used to compare MSRI-CSD approach to a conventional DTW based speaker recognition method. Section 5 presents the results and some observations. Finally section 6 presents the conclusions and future directions.

2. Capture of Speaker Dynamics in Time-Frequency Representations of Speech
From each spoken password (a 1-D signal), we create a 2-D time-frequency representation, which can be treated as an image. We resize this image to time-normalize it and then compress this image to create a 1-D representation again to obtain a fixed-dimension vector which we call CSD or compressed spectral dynamics. The essence of speaker identity, exhibited in the temporal dynamics is also exhibited in the time-frequency representation of the speech signal and is now captured by the CSD vector in an efficient and compact manner. For example, a 2-second or 16000 sample spoken password can be represented by a 35-dimension CSD vector and two spoken passwords can be easily compared by computing simply the Euclidian distance between the corresponding CSD vectors as shown in Figure 5. We have
considered two types of time-frequency representation of the speech signal here. The first one is Spectrogram (Figure 1), which is a well-known 2D temporal-spectral representation of speech, widely used for studying and analyzing the acoustic phonetic information of speech. The way a person says his/her password, i.e. the speaking style of the speaker, can be effectively captured by a well-resolved spectrogram. Spectrograms have been extensively used for forensic and legal usage of speaker recognition (but in a completely manual way) by expert human readers. Here we present an efficient & automated manner of using spectrogram for speaker recognition. Figure 1 presents a few spectrograms of the same password spoken by client speaker and an imposter. Note the within-speaker similarity and across-speaker difference and also the variable size of the x-dimension (time) due to natural variations of speaking-rate and in-between-words pause-duration.

The second time-frequency representation we used to capture speaker identity here can be viewed as a “speaker-specific-sampled” version of the spectrogram. Each speaker will have different pitch frequency (F₀ or fundamental frequency) during voiced segments and each speaker will also have different harmonic-to-noise mixture ratio. As a result, there will be speaker-specific structures in spectrograms which are efficiently captured by this newly introduced time-frequency speaker-identity feature we call “Sinogram”, which is inspired by the sinusoidal modeling of speech [12]. In a sinusoidal model, speech can be modeled by sets of sinusoids (Equation 1) whose parameters (Aₙ,θₙ,ωₙ) are estimated from the short time Fourier Transform:  

$$s(n) = \sum_{f} A_f(n) \cos(\alpha_f(n) + \theta_f)$$  

(1)

These sets of sinusoids are tracked over time using the concept of birth and death of sinusoids introduced in [12]. The proposed Sinogram is formed by placing these sinusoids on a time-frequency plane, so that the i-th sinusoid (Eqn 1) of the K-th frame becomes a point (K, Fᵢ) in Sinogram, with magnitude corresponding to Aᵢ, where Fᵢ is related to ωᵢ.

Thus Sinograms capture the natural characteristic of a person’s voice with a set of tracks of sinusoids closely related to the natural pitch and harmonic content of a person’s voice and hence closely represent speaker identity. This is somewhat evident in Figures 3 which shows the Spectrograms and Sinograms of two utterances of the client and one utterance of an imposter of the same password. It is quite evident from these examples that Sinogram preserves within-speaker similarity well while creating ample discrimination from one speaker to another.

Thus we see that the time-frequency representations, the “images” of the passwords, can be used effectively to capture speaker identity. They are quite discriminatory. However, there are two problems: a) the X-dimension is variable, making comparisons difficult, and b) the 2D data is enormous. For example, for a 2 second password or 100 frames of 20 ms each, a conventional MFCC representation will create 3900 numbers, but a 512-point DFT based time-frequency representation will mean 25600 numbers. Had we used the time-frequency representations as it is, the subsequent classifier would have an O(25.6K) complexity as opposed to a O(3.9K) complexity for MFCC. These problems are solved by the use of a suitable compression method presented next.
2.1. Compressed Spectral Dynamics Representation

We apply a suitable compression (DCT) to the time-frequency representations of the passwords, followed by a component-selection method which converts the 2D time-frequency representation to a fixed-dimension “compressed spectral dynamics” vector. DCT [14] is well known for its ability to pack the information in a small set of coefficients. DCT is also a linear, which means if \( d(AB) < d(AC) \), where \( d() \) is Euclidean distance, then \( d(DCT(A), DCT(B)) < d(DCT(A), DCT(C)) \). DCT’s compression power allows us to capture the essential information in the time-frequency representations in a small set of coefficients. We omit the DC value and keep the top \( K=(m^2-1) \) coefficients in a zigzag scan (Figure 4) to form the \( K \)-dimension CSD vector.

\[
\begin{align*}
\text{CSD} \quad \text{DCT Coefficients} \\
\end{align*}
\]

Figure 4: Formation of the CSD vector

To compare two passwords, ‘A’ and ‘B’, we thus compute two time-frequency representations, \( SGA(N \times M_1) \) and \( SG_b(N \times M_2) \), and corresponding CSDs following these steps:

a) Resize \( SGA \) to the size of \( SGB \), using a 2-D image interpolation method and form a resized \( (N \times M_1) \) \( SGR_A \).
b) Extract CSD_A and CSD_B from \( SGR_A \) and \( SGR_B \).
c) Find Euclidean distance between the two CSD’s.

Figure 5 shows the CSDs of the same Sinograms shown in Figure 3. The CSDs of the two versions of password uttered by the client is very similar, while there are noticeable differences between the CSDs of the passwords uttered by the client and the imposter. As expected, the client-to-client distance in the CSD domain (\( d=4.9 \)) is much less than client-to-imposter distance (\( d=22.6 \)), even when both client and imposter are uttering the same password.

3. Proposed Speaker Recognition Method based on Compressed Spectral Dynamics

The proposed MSRI-CSD speaker recognition method extracted a CSD vector from each password using steps described in previous section and using a simple nearest neighbor classifier for classification. We used a conventional MFCC+DTW based text-dependent speaker recognition method as baseline and compared its performance with the proposed MSRI-CSD method for different speaker recognition trials. The DTW method is based on multiple-templates as in [6] shown to deliver the best performance in its class. Proper optimization techniques in terms of selection of the right local and global constraints, as described in [11], are employed to best ensure possible performance. For a large-population speaker recognition trial a 4-template DTW will take quite a bit of time. To keep the complexity manageable, a MFCC+VQ based speaker recognition as in [9,10] is applied as a pre-selection front-end which reliably picks the best \( N \) candidates. Both the multiple-template DTW baseline and the proposed MSRI-CSD methods are applied on these top \( N \) candidates. An appropriate endpointing algorithm is used to eliminate any undue background noise/silence. Details of the multiple-template DTW baseline speaker recognition method can be found in [6]. The following approaches were taken for the MSRI-CSD speaker identification and verification trials:

**Training**: \( T \) numbers of password files are used per speaker. For DTW \( T \) sets of MFCC sequences are stored. For MSRI-CSD method, \( T \) numbers of CSD’s are extracted per speaker. A \( 8 \times 13 \) size VQ codebooks are designed using MFCC – one codebook per speaker for the pre-selection stage using the training material as well.

Given a test password, following steps were done:

1. Calculate the CSD vector \( CSD_{test} \) and (MFCC vector for DTW)
2. Pick \( N_{best} \) speakers closest (in terms of VQ distance; see [1] for details) to the test-file.
3. For each of the candidate speaker \( CP_{i} \) of the \( N_{best} \) speaker set, we calculate a distance \( D_{i} \) as follows:
   a) \( D_{i} = d(CSD_{test}, CSD_{i}) \) where \( CSD_{i} \) is the \( i \)-th stored CSD of the candidate speaker \( CP_{i} \) and \( d(X,Y) \) is a simple Euclidean distance metric. Then let \( D_{i} \) is the minimum distance over all \( T \) stored CSDs of the candidate speaker \( CP_{i} \)
   b) \( D_{o} \) is the minimum distance (DTW distance) over \( j=1,2,..,T \) templates.
   
   For DTW baseline, this distance is the DTW distance between the test MFCC sequence and the best of the stored MFCC templates for each speaker \( CP_{i} \).
   Thus for all the candidate speakers \( CP_{i} \), we create a distance array \( D = [D_{i}, D_{j}, D_{k}, \ldots, N_{best}] \), we create a distance array \( D = [D_{i}, D_{j}, D_{k}, \ldots, N_{best}] \)

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

Speaker Verification: Given a test password, and a claimed-speaker, i.e. the claimed speaker is \( i \)th in the list

1. Calculate two distance-sets (\( D_{i} \) as above): a) \( D_{o} \) using CSD as feature, and b) \( D_{VQ} \) using MFCC/VQ-distance
2. Calculate a ratio \( R_{CSD} \) and \( R_{VQ} \) as follows:
   a) \( R_{CSD} \) is one of these \( R_{CSD} = \min (D_{i}), \text{where } D_{i} = \{D_{i}, D_{j}, D_{k}, \ldots, N_{best}\} \)
   b) \( R_{VQ} \) is one of these \( R_{VQ} = \{D_{o}, D_{j}, D_{k}, \ldots, N_{best}\} \)
3. Calculate a sum and product fusion scores \( R_{sum} = 0.33 \times (R_{CSD} + R_{VQ}) \) and \( R_{prod} = (R_{CSD} \times R_{VQ})^{0.5} \times (R_{CSD} \times R_{VQ})^{0.5} \)
4. Accept if \( R_{sum} < \theta \) else reject, where \( \theta \) is some pre-computed threshold, where \( R_{sum} = \text{one of these } (R_{CSD} + R_{VQ}) \times (R_{CSD} \times R_{VQ})^{0.5} \times (R_{CSD} \times R_{VQ})^{0.5} \)

4. Experimental Set Up & Database

We needed a database in which each speaker is using unique password and each speaker is also trying to impose as a claimed speaker by saying: a) random password (unknown-password impostor trial), b) the password of the claimed 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 offer several versions of the unique client password per speaker. Therefore, we had to create our own database which to our knowledge, is the most comprehensive database for text-dependent speaker recognition (details presented in a companion paper [13] submitted to this
conference). The MSRI database has 344 speakers, recorded in an office environment over a period of 12 months. 18 of these users were recorded in multiple (4) sessions, separated by 4 weeks. Three types of unique passwords were recorded by the users: a) PWD-1: 4-digit combination (English), b) PWD-2: 4 words pass-phrase in Indian languages (mother tongue of the speaker), c) PWD-3: answers to 1 out of 10 questions, the answers being 3-5 words on average. 12-20 versions of each password are recorded by each user.

For the evaluation reported in this paper, we used a set of 246 speakers having good SNRs out of these 344 speakers and have used only PWD1 (the 4-digit password). For DTW method we used both a single template and 4 templates for training. For the MSRI-CSD method we used 4 templates. About 8-16 passwords were used for target trials for each speaker; same numbers were used for unknown-password impostor trials. For known password trials, 3-5 passwords were tried per speaker (as available in the database). This created 3309 speaker-identification trials and 8173 speaker verification trials (3309 target trials, 4181 impostor trials in which 872 trials are known-password-impostor trials). We used the following values for the system parameters: N=3 (top 3 candidates picked by pre-selection); 1st stage pre-selection VQ codebook size 16x9; MFCC dimension for DTW=39; CSD-dimension K=143; The results for the baseline DTW and the proposed MSRI-CSD system are shown in Table 1.

Table 1: Performance comparison of proposed MSRI-CSD method with conventional DTW baseline

<table>
<thead>
<tr>
<th></th>
<th>Speaker Identification Accuracy</th>
<th>Speaker Verification</th>
<th>Complexity &amp; Memory Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DTW</td>
<td>CSD T=4; K=143; CSD Fusion</td>
<td>DTW</td>
</tr>
<tr>
<td>No of MPY-ADD</td>
<td>T=1</td>
<td>T=4</td>
<td>T=4</td>
</tr>
<tr>
<td>3.2M</td>
<td>SpecG</td>
<td>SpecG</td>
<td>SpecG</td>
</tr>
<tr>
<td>12.5M</td>
<td>SinoG</td>
<td>SinoG</td>
<td>SinoG</td>
</tr>
<tr>
<td>152K</td>
<td>Sum</td>
<td>152K</td>
<td>Sum</td>
</tr>
<tr>
<td>304K</td>
<td>Pdt</td>
<td>304K</td>
<td>Pdt</td>
</tr>
<tr>
<td>3900</td>
<td>Numbers to Store</td>
<td>430</td>
<td>Numbers to Store</td>
</tr>
<tr>
<td>15.6K</td>
<td>430</td>
<td>716</td>
<td>716</td>
</tr>
</tbody>
</table>

5. Results and Observations

The results in Table 1 demonstrate that for both speaker recognition trials the MSRI-CSD method delivers competitive results compared to the baseline 4-template DTW method, while operating at significantly reduced (~1/80th) complexity with a greatly reduced (~1/35th) storage requirement. The MSRI-CSD methods deliver 0% EER for speaker verification with unknown password. Also note that among the two CSD methods, the Sinogram-CSD method delivers better results. The proposed sum and product fusion utilizing two CSDs, also deliver good results that are better than individual modes. Our experimental trials were rigorous with large variations in speakers, noise, session and content, thus the performance edge the MSRI-CSD method shows over the MFCC+DTW method do validate that for speaker recognition there is merit in processing the entire speech data than using only the spectral envelope information.

6. Conclusions & Future Directions

We presented a new way to capture speaker-identity contained in spoken passwords with various time-frequency representations and finally by a fixed compressed spectral dynamics CSD vector. The time-frequency representations proposed here offer much richer speaker-identity information than the conventional spectral-envelope-only MFCC representations. Rich speaker-specific information such as pitch, harmonics, harmonic-noise mix, are indirectly captured in the newly proposed Sinogram or a properly resolved spectrogram. The final CSD representation now represents the entire password (e.g. 2 seconds or 16000 samples) by a compact (e.g. 143 dimension) fixed-dimension vector without losing any discriminating power (as compared to 3900 numbers we need for MFCC based representation). A fixed-dimension representation of the passwords allows the use of a simple nearest-neighbor classifier. This results in performances competitive to current text-dependent speaker recognition methods at a much lower complexity. Keeping a large number of templates per speaker is also not as daunting as multi-template DTW. Another novelty of the proposed method is the capturing of the complex temporal-spectral dynamics in a speech segment as an “image”, allowing extensions like changing time-frequency resolutions as appropriate for different phonemes, noise cancellations, custom-enhancements, etc. 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, and various time-feature representations and their extensions.

7. References