16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

TEXT-DEPENDENT SPEAKER RECOGNITION BY COMPRESSED FEATURE-DYNAMICS DERIVED FROM SINUSOIDAL REPRESENTATION OF SPEECH

Amitava Das, Gokul Chittaranjan and V. Srinivasan**

Microsoft Research Lab – India; 196/36 2nd Main; Sadashivnagar; Bangalore India 560 080.
email: amitavd@microsoft.com *student interns 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 for speaker recognition based on sinusoidal representation of speech called compressed spectral dynamics (Sinogram-CSD), which effectively captures such spectral dynamics and the inherent speaker identity. The discriminative power of CSD allows classification to remain simple. The proposed CSD-MSRI method uses a simple nearest neighbor classifier to deliver performance competitive to conventional MFCC+DTW based text-dependent speaker recognition methods at significantly lower complexity.

1. INTRODUCTION

Automated speaker recognition methods can be categorized into two main types: a) text-independent (TI) and b) text-dependent (TD). Text-independent methods [1, 9-12] assume that the password users are 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. Speakers are modelled in TI methods as distributions in the feature space, captured by VQ codebooks [1,10-12] or by Gaussian mixture models [1,9]. Such distributions are often overlapping. The task of TI speaker recognition method therefore amounts to finding from which speaker distribution the test feature-vector-set is most likely to originate.

Text-dependent (TD) speaker recognition methods [2-6] on the other hand exploit the feature dynamics to capture the identity of the speaker. TD methods compare the feature vector sequence of the test utterance with the "feature-dynamics-model" of all the speakers. Such 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 utterances of the same password uttered by the speaker. For classification, conventional TD methods use dynamic classification methods such as Dynamic Time Warping (DTW) [6] or HMM [3,4].

The speaker-identity or the speaking style of a person is mostly expressed in the speech dynamics especially in the co-articulation of various sound units. TI methods miss this important aspect of speaker identity since it ignores feature dynamics treating features as a bag-of-symbols. As a result, the performance accuracies of TI methods are much lower than those of the TD methods.

Another important thing to note is that the vast majority of the prevalent speaker recognition methods, except a few [7,8], are using 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.

In this work, we propose several novel approaches for text-dependent speaker recognition. First of all we introduce a new feature called Sinogram-CSD which efficiently captures speaker identity. As we present in details later, Sinogram is a sinusoidal model based spectral representation [15] of the spoken password, which can be viewed as a speaker-specific sampling of the spectral-temporal dynamics. CSD is a novel way to represent Sinogram in a compact fixed-dimension vector. In addition to providing a high discriminating power, the Sinogram-CSD feature provides the unique simplicity of representing variable-length spoken passwords by fixed-dimension CSD vectors. This enabled us to build a CSD-MSRI TD speaker recognition method which does not require any highly-complex dynamic classification methods such as DTW or HMM. The CSD-MSRI method deploys a simple nearest-neighbor classifier to deliver high performance, competitive to conventional TD methods, at a significantly lower complexity. Our proposed approach therefore looks well beyond the envelope-only features of conventional methods and utilizes the entire speech signal.

The paper is organized as follows: Section 2 presents the new Sinogram-CSD feature. Section 3 presents the proposed CSD-MSRI TD speaker recognition method. Section 4 presents the database, experimental details and results. Finally section 5 presents the conclusions and future directions.

2. SINOGRAM: SPEAKER-RECOGNITION FEATURE DERIVED FROM SINUSOIDAL REPRESENTATION 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 are 2D temporal-spectral representation of speech. Spectrograms have been a highly useful and popular means of studying and analyzing the acoustic phonetic information in speech signal and today spectrograms are quite extensively used for forensic and legal usage of speaker recognition. There are a number of experts who are trained to read spectrograms to identify the content as well as the speaker. We propose a novel feature, suitable for speaker recognition, that we call the "Sinogram" (Figure 1) which can be viewed as a "speaker-specific-sampled" version of the spectrogram. Like Spectrogram, Sinograms also capture the temporal-spectral variations in speech, but in a speaker-specific manner, which makes it so effective for TD speaker recognition.

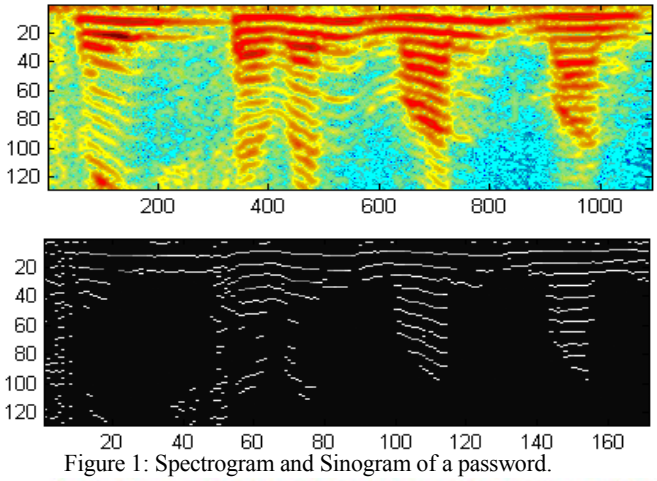

Figure 1: Spectrogram and Sinogram of a password.

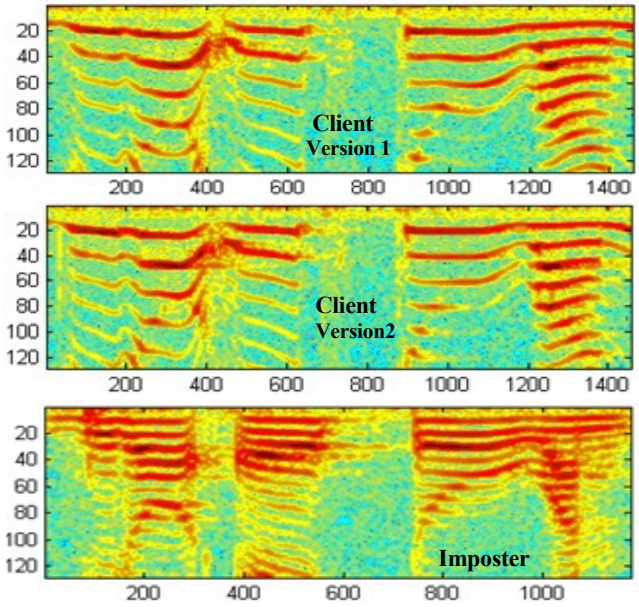

Figure 2 Spectrograms of passwords spoken by the client (top two) and an imposter (bottom one) uttering the same password

The concept of Sinogram is inspired by sinusoidal modelling of speech [15] which has been successfully used in speech coding and speech manipulation such as voice transformation in the past. Essentially, speech can be modelled by sets of sinusoids whose parameters are estimated from the short time Fourier Transform:

$$s(n) = \sum_{l=1}^{L(n)} A_l(n) \cos(\omega_l(n) + \theta_l) \quad \ldots (1)$$

These sets of sinusoids are tracked over time using the concept of birth and death of sinusoids as introduced in [15]. Thus, the number of sinusoids that appear from frame to frame is variable. The proposed Sinogram is formed by placing these sinusoids on a time-frequency plane, so that the l-th sinusoid (Equation 1) of the K-th frame becomes a point (K, F_l) having magnitude corresponding to A_l. The F_l index corresponds to the frequency ω_l.

Therefore Sinograms capture the natural characteristic of a person's voice with a set of tracks of sinusoids (Figure 1 & 3). The characteristics of these tracks are 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 1-3. Figure 1 shows Sinograms as a simplified view of the spectrogram, but highly speaker-oriented. Note how spectrograms capture speaker identity as shown in an example in Figure 2, showing the spectrograms of the same password spoken by the client (different two times) and spoken by an imposter. The within-speaker similarity and speaker-to-speaker discrimination is quite evident. Figure 1 and Figure 3 show that, like spectrogram, speaker identity can be captured efficiently by a properly resolved Sinogram. It is quite evident from the example in Figure 3 that Sinogram preserves within-speaker similarity well while creating ample discrimination from one speaker to another.

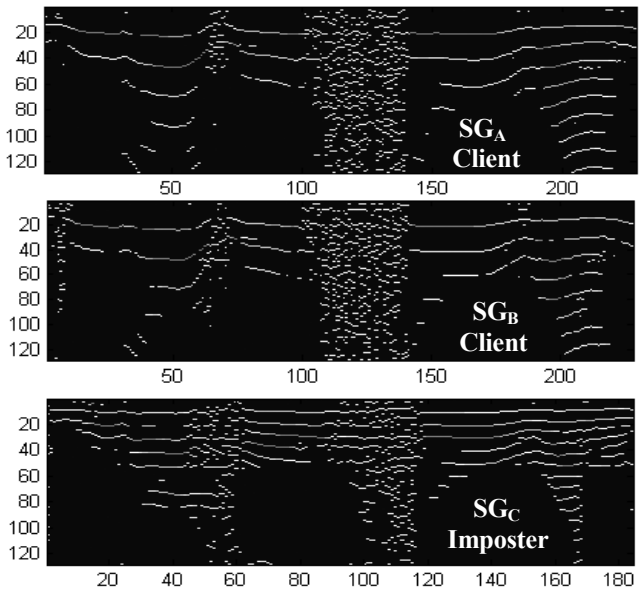

Figure 3: Sinograms corresponding to the passwords in Figure 2 (top two uttered by the client, bottom one spoken by an imposter)

However, the Sinogram representation introduced here has two major barriers: a) variable dimensionality and b) large data size. If a password has M number of segments of length T samples (typically T =160-256) and each segment is converted to a set of N frequency points, this creates a NxM 2D Sinogram (M, the x-dimension will vary from one password to another). For example a password of 100 frames resolved into a 128 point Sinogram will amount to storage of 12800 numbers (compared to 3900 numbers for an MFCC+delta+double-delta representation). To compare two variable-dimension Sinograms of size N1xM and N2xM, dynamic programming methods like DTW can be applied but the process will be incredibly complex.

These two problems are solved by the Compressed Spectral Dynamics (CSD) feature introduced next.

2.1 CSD: Compressed Sinogram Signature

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

Speech password → *Sinogram* ***SG*** → *Apply Transform, W to obtain* ***T****= W(S)* → ***CSD*** *= f_select(T)*

Where W is a suitable transform and "f_select" is a suitable function which picks K components from T (see Figure 4)

Examples of CSDs of different passwords are shown in figure 5. Note that now the entire speech utterance is represented by a fixed K-dimension CSD vector.

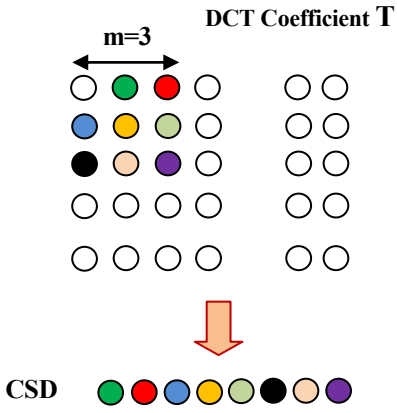

Figure 4: Formation of CSD vector from DCT Coefficients T

To compare two passwords, 'A' and 'B' , we compute two Sinograms, SG_A(NxM1) and SG_B(NxM2), and corresponding CSDs following these steps:

a) Resize SG_B to the size of SG_A using a 2-D image interpolation method and form a resized Sinogram $SG_{B'}$

b) Extract CSD_A and $CSD_{B'}$ from $SG_{B'}$ (NxM1) and SG_A

c) Find distance between the two CSD's as shown below.

dist(utterance-A, utterance-B) ~= dist(SG_A, SG_B) ~= dist(SG_A, $SG_{B'}$) ~= dist(CSD_A, $CSD_{B'}$)

A suitable choice of transform ensures that the crucial distance property (shown below; 'dist' can be a simple Euclidean distance) is maintained:

dist(A,B) < dist(A,C) → *dist(CSD_A, CSD_B) < dist(CSD_A, CSD_C),*

We have chosen Discrete Cosine Transform [13] as the transform which preserves the above distance property as well as packs most of the information in a small sets of transform coefficients allowing us to define a "f_select" function as illustrated in Figure 3 below. We omit the DC value and keep the top K=(m^2-1) coefficients (see Figure 4) in a zigzag scan to create the K-dimension CSD vector.

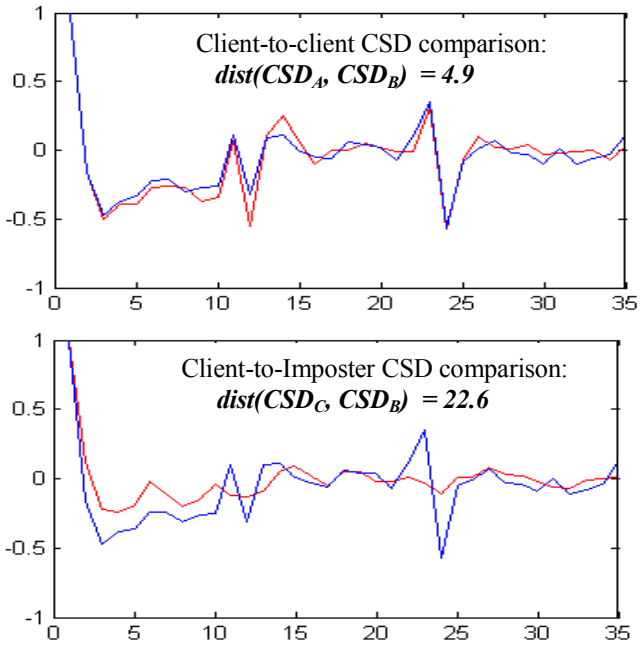

Figure 5: CSD representation of client and imposter passwords derived from Sinograms shown in Figure 3. Top Figure compares the CSDs from the two passwords from the client (CSDs derived from the top two Sinograms of Figure 3) and the bottom figure compares the CSDs of the client and imposter passwords (CSDs derived from the bottom two Sinograms of Figure 3)

Figure 5 illustrates a few important things. First of all, the "variable X dimension problem" of Sinograms is gone as they are converted into fixed dimension CSD vectors (m=6; K=35 in this example. Furthermore, the within-speaker similarity and across speaker discrimination of Sinogram (as observed in Figure 3) are left intact in the CSD domain, as evident from the figures and distances presented in Figure 5.

Thus Sinograms and CSDs do capture speaker identity well by preserving the speaker-specific temporal-spectral dynamics in the speech signal. Also note that a conventional MFCC+DTW method needs 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 significantly compared to conventional TD methods.

3. MFCC-VQ AND SINOGRAM-CSD BASED TWO-STAGE SPEAKER RECOGNITION METHOD

We present the details of the proposed MSRI-CSD speaker recognition algorithm here. The MSRI-CSD method has an MFCC+VQ based static classification 1st stage which picks N_{best} 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 N_{best} candidates using CSD and makes the final decision. The key algorithms of the two stages are given below.

First Stage Algorithm: Design S CxD size VQ codebooks $\{\mathbf{CB_1}, \mathbf{CB_2}, \ldots \ldots \mathbf{CB_S}\}$, 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 $\{\mathbf{F_1}, \mathbf{F_2}, \mathbf{F_m} \ldots \mathbf{F_M}\}$ and follow these steps:

Step-1: *For each speaker P_i compute an accumulated distance A_i as follows: $A_i(m) = D_i(1) + D_i(2) + \ldots\ldots + D_i(m)$ where $D_i(m)$ is the minimum distance of $\boldsymbol{F_m}$ to $\boldsymbol{CB_i}$ (codebook of person P_i); D_i = minimum of D_{ij}, where $D_{ij} = \|\boldsymbol{F} - \boldsymbol{C_{ij}}\|^2$ $j=1,2,\ldots,L$, $\boldsymbol{C_{ij}}$ being the code vector of the codebook $\boldsymbol{CB_i}$*

Step-2: *Pick the N_{best} speakers, for which $A_i(M)$ are lowest, i.e. for N_{best} =3, pick the 3 speakers having the 3 lowest $A_i(M)$ scores*

After 1st stage we have a selected set of candidates CP =[CP_i, i=1, 2,3, … N_{best}]

Second Stage Algorithm: For each of the candidate speaker CP_i, in the N_{best} speaker set, we calculate a distance D_i as follows:

Let D_{ij} = dist(SG_test, SG_template(i,j))) = d(CSD_test, CSD(i,j)), where SG_test and CSD_test are the sinogram and CSD of the test utterance and SG_template(i,j) and CSD(i,j) are the sinogram and CSD of the j-th stored password template of the candidate speaker CP_i;

Let D_i be the minimum distance over all D_{ij} over j=1,2..P templates. Thus for all the candidate speakers CP =[CP_i, i=1, 2,3, … N_{best}] we create a distance array D = [D_k, k=1, 2,3, … N_{best}]

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

For Speaker Verification:

1. Create a modified candidate list CP'=[target-speaker, N_{best} speakers excluding target-speaker], i.e. the target speaker is 1st in the list.

2. Calculate two distance-sets (D=[D_k] as above): a) $\mathbf{D_{CSD}}$ using CSD as feature, and b) $\mathbf{D_{VQ}}$ using MFCC-VQ-distance $A_i(M)$

3. Calculate two likelihood ratios $\mathbf{R_{CSD}}$ & $\mathbf{R_{VQ}}$ as follows R= d1/d2, where d2=D(1); d1= minimum(D'), where D'=[D(2) … D(N_{best})]

4. *Calculate a product fusion score* $\boldsymbol{R_p} = \mathbf{R_{CSD}} \times \mathbf{R_{VQ}}$

5. Accept if $\boldsymbol{R_p} > \theta$ else reject, where θ is some pre-computed threshold

Therefore, the CSD-MSRI method employs both static and dynamic classifications and judiciously utilizes envelope (MFCC) as well as entire speech signal information.

The system parameters of the CSD-MSRI method are: C=size of 1st stage VQ codebook, D=dimension of 1st stage VQ codebooks, N_{best}, K=Dimension of CSD and P=number of password templates per speaker. Note that the proper parameters (window-size, hop-size and DFT-size) should be chosen to create an appropriately-resolved Sinogram.

4. DATABASE, EXPERIMENTAL SETUP, BASELINE SPEAKER RECOGNITION SYSTEM & RESULTS

The CSD-MSRI method is compared with a MFCC+DTW speaker recognition system (similar to [6]) which uses 39-dim MFCC and conventional DTW[6,14] with simplest local paths and r=8 as global constraint. We have not tried any optimization techniques for the baseline DTW. We used a single template per speaker for CSD-MSRI and 1 and 4 templates per speaker for the DTW baseline. To give the same treatment, the same MFCC-VQ 1st stage is applied to the DTW-Baseline method as well.

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 saying: a) random password (unknown-password imposter trial) , b) the password of the target speaker (known-password imposter 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 the unique client password, i.e. several unique client passwords per speaker. Therefore, we for last one year, we have been collecting data to create a unique MSRI Speaker Recognition database. To our knowledge, this is the most comprehensive database (publicly available from MSRI for research usage) for text-dependent speaker recognition tasks.

The MSRI database has 370 speakers, recorded in an office environment over a period of 12 months. 20 of these users were recorded in multiple 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 2 to 60 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 260 speakers; used 4 passwords for training and remaining ones for testing. This created 3309 speaker-identification trials and 8173 speaker verification trials (3309 target trials, 4181 imposter trials in which 872 trials are known-password-imposter trials). The results for the baseline DTW and the proposed CSD-MSRI system are shown in Table 1.

Performance Comparisons		DTW-1 template	DTW-4 templates	CSD-MSRI 1 template
SID results in % Error		12.00%	5.01 %	0.17 %
SV results in % EER	unknown PWD	4.11%	0.38 %	0.00 %
	known PWD	9.01%	4.90 %	0.10 %
Complexity Comparisons				
Storage: numbers to store/speaker		3900	15600	287
Complexity in terms of MPY-ADD		3120K	12480K	38K

Table 1: Performance & complexity comparisons of the CSD-MSRI with MFCC-DTW [Assumptions:100 frames on average per password. Other Data: 260 speaker recognition task; N_{best}=3; and C=16;D=9; P=1; K=143 for CSD-MSRI method (section-3)]

5. CONCLUSIONS AND FUTURE DIRECTIONS

As seen in Table 1, the proposed Sinogram-CSD 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 quite competitive and the complexity is significantly lower (storage complexity: a factor of ~50 less; operational complexity: a factor of 300 times less, both w.r.t. 4-template DTW). Both identification and verification tasks were well handled by the CSD-MSRI method.

Note that, the speaker verification trials in which imposters do not know target passwords (unknown-PWD) are relatively easier than the known-PWD trials, as in the former the password-discrimination can be exploited. As a result, the DTW performances are better, but the CSD-MSRI exploited the password discrimination well demonstrating 0% EER or no errors, i.e. completely separated client-imposter score distributions. When we randomly picked various speaker-sets and trials, we have observed similar results (zero errors or almost no errors) for CSD-MSRI.

For known-PWD case, since MFCC largely represent the 'content' or the 'text' in speech, and since the content is same here in passwords uttered by imposters and clients, the conventional MFCC+DTW method could not perform that well and confused imposters with clients in many cases. On the other hand, the CSD-MSRI method shows better results mainly due to the inherently better speaker-discrimination power of the Sinogram-CSD feature. Our experimental trials were rigorous with large variations in speakers, noise, session and content. Thus, the performance edge the CSD-MSRI method shown here 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. The newly proposed Sinogram-CSD feature looks quite promising for speaker recognition.

To summarize, we introduced a sinusoidal analysis based new speaker-recognition feature called Sinogram-CSD which also provides a compact fixed dimension vector representation of the entire speech password enabling simpler classification and lower complexity. The proposed CSD-MSRI TD speaker recognition method demonstrated competitive performance over conventional TD methods at a significantly reduced complexity. This opens a promising new horizon for speaker recognition research and at present we are pursuing this further with different normalization and transformation techniques and various other speech models.

6. REFERENCES

[1] Bimbot et al, "A Tutorial on Text-Independent Speaker Verification", Eurasip J. Appl. Speech Proc. 4 (2004)

[2] A. Das & V. Ram, "Text-dependent speaker-recognition – A survey and State of the Art", Tutorial at ICASSP-2006

[3] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," Proc. ICASSP-93

[4] D. Falavigna, "Comparison Of Different HMM Based Methods For Speaker Verification", Proc. Eurospeech-95

[5] Higgins et al., "Speaker Verification using randomized phrase prompting", Digital Signal Processing, 1(2):(1991)

[6] V. Ram, A. Das, and V. Kumar, "Text-dependent speaker-recognition using one-pass dynamic programming", *Proc. ICASSP'06*, (2006)

[7] Murty & Yegna, "Combining Evidence From Residual Phase and MFCC Features for Speaker Recognition", IEEE Signal Processing Letters,V13-1, (2006).

[8] Zheng et. al., "Integration of Complementary Acoustic Features for Speaker Recognition", IEEE Signal Processing Letters,V14-3, (2007).

[9] D. Reynolds, et al, "Speaker Verification using adapted GMM", Digital Signal Processing, 10(1-3), 2000.

[10] F.K.Soong, et al, "A vector quantization approach to speaker recognition", *AT&T Tech. J.,* Vol 66 (1987)

[11] A. Das & P. Ghosh, "Audio-Visual Biometric Recognition by Vector Quantization", Proc. *IEEE SLT-06*

[12] T. Kinnunen, E. Karpov, and P. Franti, "Real-Time Speaker Identification and Verification", *IEEE Trans. On Audio, Speech and Language Processing*, V14-1, Jan 2006.

[13] K. Rao & P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications* (Academic Press (1990).

[14] Sakoe & Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Tran* ASSP-26-1 (1978)

[15] McAulay & Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE ASSP, 1986.

.