DIRECT MODELING OF SPOKEN PASSWORDS FOR TEXT-DEPENDENT SPEAKER RECOGNITION BY COMPRESSED TIME-FEATURE REPRESENTATIONS

Amitava Das and Makarand Tapaswi *

Microsoft Research, Redmond, WA *intern at MSRI
amitavd@microsoft.com

ABSTRACT

Traditional Text-Dependent Speaker Recognition (TDSR) systems model the user-specific spoken passwords with frame-based features such as MFCC and use DTW or HMM type classifiers to handle the variable length of the feature vector sequence. In this paper, we explore a direct modeling of the entire spoken password by a fixed-dimension vector called Compressed Feature Dynamics or CFD. Instead of the usual frame-by-frame feature extraction, the entire password utterance is first modeled by a 2-D Featurogram or FGRAM, which efficiently captures speaker-identity-specific speech dynamics. CFDs are compressed and approximated version of the FGRAMs and their fixed dimension allows the use of simpler classifiers. Overall, the proposed FGRAM-CFD framework provides an efficient and direct model to capture the speaker-identity information well for a TDSR system. As demonstrated in trials on a 344-speaker database, compared to traditional MFCC-based TDSR systems, the FGRAM-CFD framework shows quite encouraging performance at significantly lower complexity.

Index Terms— Speech Features, Text-Dependent Speaker Recognition

1. INTRODUCTION

Conventional Text-Dependent Speaker Recognition (TDSR) systems [2] use a unique password for each user and from the spoken password derive the user-identity by typically extracting frame-by-frame spectral features like MFCC[1,2]. Due to the natural variations of the speaking-rate, even if the same speaker says the same password twice, the length of the feature-vector sequence varies from one password to another. To compare two such variable-length spoken passwords, conventional TDSR methods employ dynamic classification techniques, such as DTW [9] or HMM [4]. DTW based TD systems capture the speaker-specific speech dynamics information as multiple raw templates of the feature-vector sequences extracted from training data and excellent results were reported for TDSR applications in [9]. HMM based TD methods capture the speaker-specific speech dynamics as adapted HMMs, one for each user. Excellent results have also been reported in the past using these schemes [4, 5, 8]. However, both DTW and HMM based TDSR methods, require a good amount of storage and computational complexity. Some of the HMM based systems [6, 8] also require a full-blown ASR engine as part of the system. In all of these conventional TDSR systems, a frame-by-frame extraction of features and spectral-envelope only features, such as MFCC, have been used.

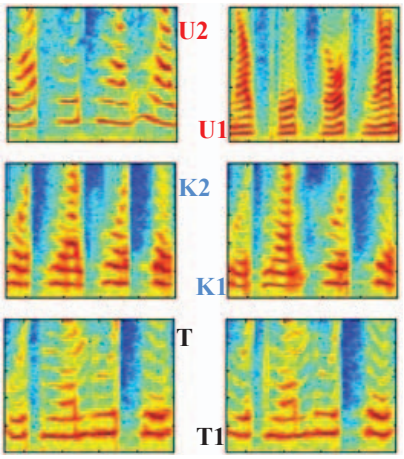
In this paper, we explore two main ideas. First, we propose that for text-dependent speaker recognition where each speaker is using a unique password, it is better to consider the entire password as a whole entity as opposed to looking at it frame-by-frame which makes it too granular. We propose a direct model of the spoken password using a framework we call FGRAM-CFD. In this framework, the entire password utterance is first represented by a 2-D time-feature representation we call "Featurogram" or FGRAM, which is like an "image" of the speaker-identity-information. FGRAM is found to be a powerful feature for speaker recognition as shown in later sections. CFDs are compressed versions of FGRAMs which even though compressed do retain the discriminatory power of the FGRAMs. Thus with the proposed FGRAM-CFD direct model, each password, irrespective of the length of utterance, is now represented by a fixed dimension CFD feature vector. For example, a 3 second long password will be represented by a vector of size 143 (as compared to 39x300 numbers in conventional MFCC+Delta feature representation). This reduces storage complexity but more importantly this new CFD feature allows a TDSR system to use simple classifiers as the variable-dimensionality problem is solved. To compare two passwords of different length we no-longer need any dynamic programming type methods such as DTW [9]. As a result, overall storage and computational requirements are greatly reduced.

The second key suggestion of our paper is to look beyond features specific to the spectral-envelope (e.g. MFCC) and to use the entire signal content. Speech production is modeled as a convolution of an excitation signal with a spectral-envelop or glottal-shape function. Thus, by using spectral-envelop-type features such as MFCC, we are not utilizing a lot of speaker-specific information contained in the excitation part, such as pitch, harmonics, extent of voicing, etc. In this paper, we explore a properly resolved spectrogram as an FGRAM feature which represents the entire signal. The results show that looking at information beyond spectral-envelope indeed does help as spectrogram-CFD performs better than MFCCgram-CFD. Combining multiple types of FGRAMs also helps to enhance performance. It also helps to combine the frame-by-frame MFCC feature and the direct-model CFD feature as seen in the hybrid FGRAM-CFD framework proposed

here for a TDSR system we are building for in-office log-in/access control type applications. Our paper is organized as follows: Section 2 presents the proposed FGRAM-CFD approach and section 3 presents the multi-CFD speaker recognition method. Section 4 details the experimental comparisons with conventional TDSR methods. Section 5 presents the results. Finally section 6 presents the conclusions and future directions.

2. FGRAM-CFD: A DIRECT MODEL OF SPOKEN PASSWORD FOR SPEAKER RECOGNITION

We believe that the speaking style of a person is embedded in the temporal dynamics of the speech feature and therefore for speaker recognition it is beneficial to capture the complete time-feature dynamics of the entire password. The proposed FGRAM feature does capture the feature dynamics of the whole password and captures the speaker-identity quite well. This is illustrated in Figure 1 which shows the spectrograms of passwords spoken by the client speaker and imposters in a TDSR system. Note the within-speaker similarity (client case) and across-speaker divergence (for both known and unknown password cases).



Unknown-PWD-imp case Imposters saying PWD which is different than the client speaker

Known-PWD-imp case Imposters saying PWD which is the SAME as the client speaker

Client case client speaker saying his/her PWD at two different time instants

Figure 1: Spectrogram of same passwords spoken by client-speaker at two different times (T & T1 – client-case); The same password spoken by imposters (K1 & K2 – known-password imposter case); Passwords (different than client) spoken by other imposters (U1, U2: unknown-password-imposter case - imposter guessing password of the client.).

As seen in figure 2, an N1xK size FGRAM is formed by simply stacking the N1, K-dimensional feature-vectors extracted from the N1 frames of the password. In this paper, we explore the use of two types of FGRAMs namely a) Spectrogram and b) MFCC-FGRAM, formed by stacking K-dimension MFCCs. Clearly the 2-D FGRAM feature does capture a lot of information about the speaker-identity by taking a "snap-shot" of the entire password. But an FGRAM would require lots of numbers to store. Also the variable-dimensionality of the X-axis of the FGRAMs remains a problem.

Both these problems are solved by the 2nd step of our FGRAM-CFD framework as we convert each FGRAM image to a compressed and fixed-dimension CFD vector by applying a 2-D Discrete Cosine Transform to the FGRAM followed by a specific truncation of the DC-coefficient and the higher order DCT coefficients as shown in Figure 2.

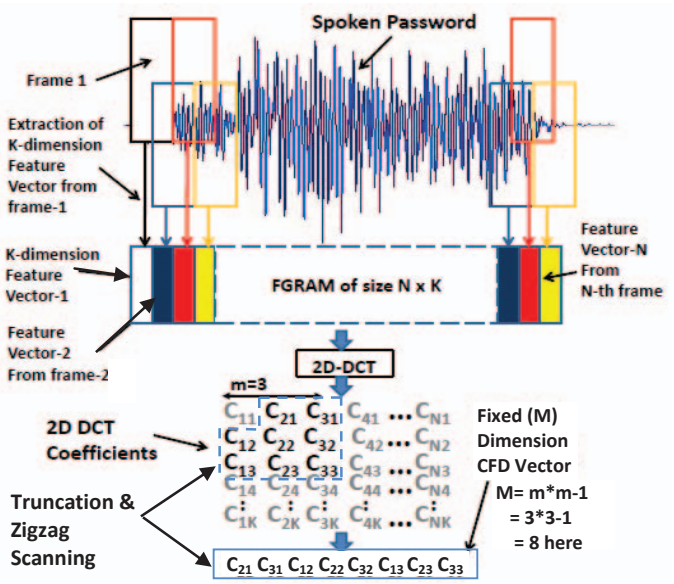
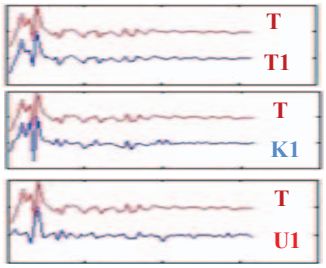


Figure 2: Formation of FGRAM & FGRAM-CFD from a spoken password

Thus the proposed FGRAM-CFD framework directly models the spoken passwords by this M-dimension CFD vector. Even though the information is compressed, CFD retains the discriminative power of the original FGRAM as shown in Figure 3 below. The "visual" discrimination evident in the FGRAMs of Figure 1 is retained in the CFDs as shown by the quantitative distances in Figure 3.



Client-to-client comparison
dist(CFD_T, CFD_T1) = 0.089

Client-to-Imposter comparison
Known-PWD-imp Case
dist(CFD_T, CFD_K1) = 0.744

Client-to-Imposter comparison
Unknown-PWD-imp Case
dist(CFD_T, CFD_U1) = 3.254

Figure 3: CFD comparisons (target-to-target and target-to-imposter) for the same password-spectrograms shown in Fig 1.

Overall, the FGRAM-CFD direct model offers the following benefits: a) it captures speaker-identity information contained in the spoken password in a more holistic manner than frame-by-frame features which are too granular, b) it offers a compact and fixed-dimension representation of the spoken password, c) it allows easy comparison of two passwords and therefore allows the use of simple classifiers, and d) it provides an easy "integration" of any type of new speakerID features as in [7] to this FGRAM-CFD framework.

3. A TWO-STAGE HYBRID MFCC+MULTICFD FRAMEWORK FOR TD SPEAKER RECOGNITION

For our main objective of building an efficient TDSR system for office login/access-control, we propose a nearest-neighbor classifier based framework which utilizes all available information in the spoken password judiciously, namely frame-by-frame MFCC feature as well as the direct model FGRAM-CFD features using multiple types of FGRAMs. The key steps are described next:

System Parameters: *U: total number of users; F: number of different types of FGRAMs used; K: dimension of per-frame MFCC feature; M: dimension of CFD; T: no of password templates for training; L: MFCC-VQ CB size;*

Training: *For each user P_u, from the T training templates, design F CFD codebooks $\boldsymbol{CCB^k_u}$ of size TxM, k=1,2,…,F; one for each of the F types of FGRAMs. Also use conventional VQ methods [10] to design a MFCC VQ codebook $\boldsymbol{MCB_u}$ of size LxK from MFCCs extracted from all T password training-templates of user P_u.*

Test: *Given a test password having N1 frames, generate MFCC vectors $\boldsymbol{V_j}$, j=1,2,…,N1 and F FGRAM-CFDs $\boldsymbol{W_k}$, k=1,2,…,F. Given an identity-claim of "c" (i.e. it is spoken by user P_c), compute two inverse-likelihood-ratios $\mathbf{R_{CFD}}$ and $\mathbf{R_{MVQ}}$ as follows:*

$\mathbf{R_{MVQ}} = \sum_{j=1}^{N1} \boldsymbol{D_{TGT}}(j) / \sum_{j=1}^{N1} \boldsymbol{D_{NEXT}}(j)$; $\boldsymbol{D_{TGT}}$ *and* $\boldsymbol{D_{NEXT}}$ *being:*

$\boldsymbol{D_{TGT}}$ **(j)** *is the minimum distance of $\boldsymbol{V_j}$ from MFCC codebook $\boldsymbol{MCB_c}$ of claimed person P_c, $D_{TGT}(j) = \min\{D_m\}$, $D_m = ||\boldsymbol{V_j} - \boldsymbol{MCB_{cm}}||^2$, m=1,2,…,L.*

$\boldsymbol{D_{NEXT}}$ **(j)** *is the minimum distance of $\boldsymbol{V_j}$ from the set of all codebooks of all other users except P_c , $D_{NEXT}(j) = \min\{D_{mn}\}$; $D_{mn} = ||\boldsymbol{V_j} - \boldsymbol{MCB_{mn}}||^2$, m=1,2,…,L; n=1,2,..,U; $n \neq c$.*

For each of the F types of FGRAMs, the $\mathbf{R_{CFD}}(k)$ ratio is computed from the corresponding CFD $\boldsymbol{W_k}$ as follows:

$\mathbf{R_{CFD}}(k) = \boldsymbol{D^k_{TGTCFD}} / \boldsymbol{D^k_{NEXTCFD}}$; *where D^k_{TGTCFD} is the minimum distance of the test CFD vector $\boldsymbol{W_k}$ from CFD codebook $\boldsymbol{CCB^k_c}$ of the claimed person P_c, $D^k_{TGTCFD} = \min\{D^k_m\}$, $D^k_m = ||\boldsymbol{W_k} - \boldsymbol{CCB^k_{cm}}||^2$, m=1,2,…,T, and $\boldsymbol{D_{NEXTCFD}}$ is computed in the same manner as $\boldsymbol{D_{NEXT}}$ above, but using CFD codebooks.*

Then either SUM or PRODUCT fusion are used to combine the $\mathbf{R_{CFD}}(k)$ scores of the F different FGRAMs to form the final $\boldsymbol{R_{CFDF}}$ score as:

$$\mathbf{R_{CFDF}} = \left(\frac{1}{F}\right) \sum_{k=1}^{F} \boldsymbol{R}_{\mathrm{CFD}}(k) \quad \text{or} \quad \mathbf{R_{CFDF}} = \prod_{k=1}^{F} \boldsymbol{R}_{\mathrm{CFD}}(k)$$

The final scoring ratio $\boldsymbol{R_f}$ is computed as $\boldsymbol{R_f} = \boldsymbol{R_{MVQ}} * \boldsymbol{R_D}$ where $\boldsymbol{R_D}$ is defined as:

$\boldsymbol{R_D} = 1$ *if* $\boldsymbol{R_{MVQ}} < \theta_L$ *or* $\boldsymbol{R_{MVQ}} > \theta_H$

$= \boldsymbol{R_{CFDF}}$ *otherwise.* $[\theta_L, \theta_H]$*: pre-determined constants.*

The final decision is made as follows: For speaker verification, we compare $\boldsymbol{R_f}$ with a threshold λ and accept the claim if $\boldsymbol{R_f} < \lambda$ and reject otherwise. For speaker identification, the user for which $\boldsymbol{R_f} < 1$ is chosen as the identified user. This way, both the frame-by-frame MFCC feature and the direct-model CFD feature from the entire password are utilized. Note that the 1st stage MFCC-VQ is used as a pre-selection step which enhances the performance.

4. EXPERIMENTAL DETAILS AND DATABASE

For our research, we needed a TDSR database having a large number of "client" speakers saying their unique passwords several times as well as many "imposter" speakers saying random passwords (unknown password imposter) as well as passwords of other clients (known-password imposter). We could not find any such publicly-available database. The closest one is LDC-YOHO but it does not offer several versions of the unique client password. Therefore, for this research and for our objective to build a TDSR system for office log-in/access-control we created our own TDSR database having 344 speakers recorded in realistic office environment with usual office noise and SNR conditions (i.e. not a clean studio recording) over a period of 9 months.

This MSR TDSR database is publicly available (please contact the main author) for research purpose and we encourage the speaker recognition community to use it as it is really a good database for Text-Dependent Speaker Recognition research with ample examples of realistic client and imposter (both known and unknown-password) passwords in actual office environment, containing for each speaker 15-20 utterances of his/her unique password, 0-8 known-passwords of other speakers, and 6-10 random passwords. More details of this TDSR database can be found in [11].

For baseline comparison we used a password-HMM TDSR system as in [4] using 39 dimension MFCC+Delta feature. Several system combinations were tried and the combination of 12 password templates for training, 12 states and 4 mixtures per state gave the best performance and this combination is used as the baseline. We also used the DTW TDSR system as in [9] as baseline which reported excellent results using dynamic programming with multiple password templates per user. Same MFCC+Delta feature was used and 6 password templates were used for training. As the proposed framework is for a TDSR system, we have not included any TI (text independent) methods such as GMM [1] or GSV as baseline. Usually TD methods [2] deliver better performance than TI methods. TI methods also do not perform well for test utterance of short duration like our typical 2-second long passwords. We did not implement the HMM system in [8] yet but plan to do so in near future.

For the proposed CFD system, we have chosen the system parameters as: No. of password templates for training: T=4; $[\theta_L, \theta_H] = [0.95, 1.1]$; VQ CB size=8x9; CFD-dimension M is chosen as 143 (see Table 1 which shows the impact of M on the speakerID performance). An appropriate "end-pointing" (EP) method removes the silence/noise part of the spoken passwords before forming the FGRAM.

All system conditions including the cohort normalization, end-pointing, the 1st-stage VQ-gating are kept the same for the baseline systems. Only differences are that the baseline systems used more training templates and the conventional MFCC+Delta feature. All passwords other than those used for training are used for testing. This created a total 1573 identification trials and 4257 verification trials (1573 target and 2684 imposter trials in which 1111 are known-password-imposter trials).

5. RESULTS AND DISCUSSIONS

Table 1 shows the impact of the CFD dimension M on performance, Table 2 presents the impact of the various combinations of the CFD framework and Table 3 compares the CFD framework with the baseline methods. In Table 2, the "CFD-alone" row indicates a condition when the 1st-stage VQ pre-selection is not used. The "two-stage" row represents the proposed two-stage framework which works better than the "CFD-alone" condition. Also we see that spectrogram-CFD performs better than MFCCgram-CFD, i.e. for speaker recognition it is beneficial to include information from the entire speech signal as compared to using only spectral-envelope type information. Sum fusion of the two FGRAMs gives the best result for both identification and verification tasks. Note that the speaker verification results reported here are for the tougher "known-password" condition. For the unknown-password condition, the results are significantly better.

M	15	35	63	99	**143**	224
SID Err. (%)	22.83	10.67	9.33	7.67	**5.67**	6.17

Table 1: Impact of CFD dimension M on Speaker ID performance (100 speaker; 4 templates/speaker; 600 trials; Spectrogram as FGRAM)

		Spectro gramCFD	Mfcc gramCFD	Product Fusion	Sum Fusion
SID %Error	CFD-alone	5.9	6.2	0.121	0.112
	Two-Stage	1.2	1.6	0.110	0.001
SV %EER	CFD-alone	3.8	4.7	1.651	1.550
	Two-Stage	0.92	0.95	0.420	0.108

Table 2: Performance of various combinations of proposed CFD system

Comparisons	DTW	HMM	FGRAM-CFD
SID performance (%Error)	0.93	0.32	0.001
SV performance (%EER)	4.9	2.3	0.108
Storage (no. to store)	15.6K	3.8K	600
Complexity (no. MPY-ADD)	12.5M	1.2M	300K

Table 3: Performance and complexity comparisons of proposed FGRAM-CFD method (two-stage-with-sum-fusion) with baseline methods

As seen in Table 3, the proposed FGRAM-CFD framework offers encouraging performance compared to conventional TDSR baseline at a significantly lower complexity. Especially for the tougher known-password speaker verification condition, the CFD method performs better. Note that the CFD method used 4 password templates per user for training, while the DTW and HMM methods used 6 and 12 respectively. We have kept it low for the CFD system because in reality people do not want to "record" many password templates during enrollment.

6. CONCLUSION AND FUTURE DIRECTIONS

We presented a new and interesting approach to text-dependent speaker recognition by directly modeling the spoken password by a fixed dimension FGRAM-CFD vector. In contrast to traditional frame-by-frame processing, (which we feel is too granular) the proposed FGRAM-CFD framework enables us to look at broader speaker-specific speech dynamics contained in the entire span of the spoken password. The CFD vector, though an approximation, still retains the discriminating power of the 2-D FGRAMs. Fixed dimensionality of CFD allows one to use simpler classifiers, as there is no more any need to use complex dynamic programming method like DTW to handle the variable dimensionality factor (as needed in traditional frame-by-frame MFCC type feature based systems). This creates a powerful and discriminatory model for TDSR while keeping the storage and computational complexity low.

We also explored spectrogram as an FGRAM and found that it works better than MFCCgram. This clearly shows that is useful to look beyond the presently-popular spectral-envelop-only information as MFCC and incorporate more information from the entire speech signal. Experimental evaluation on a large 344 speaker TDSR database had shown that the proposed FGRAM-CFD TDSR framework delivers quite encouraging performance, as good as conventional TDSR methods, while using significantly less computational and storage requirements.

The FGRAM-CFD paradigm detailed here presents an interesting approach to model variable-length speech segments of interest and thus we are exploring the use of the FGRAM-CFD model for other speech applications as well.

7. REFERENCES

[1] Bimbot, Reynolds, 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 Presented at ICASSP-2006.

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

[4] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," Proc. ICASSP-93, pp.391-394.

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

[6] Qi Li, B. Juang, Q. Zhou, C-H Lee, "Automatic Verbal Information Verification for User Authentication", IEEE Tran-SAP, vol 8, no5, Sep-02

[7] Zheng Z. et. Al., "Integration of Complementary Acoustic Features for Speaker Recognition", IEEE SP Letters,V14-3, (2007)

[8] Subramanyal, A. Zheng Z., et al, "A Generative Framework using Ensemple Methods for Text-Dependent Speaker Verification", Proc. ICASSP-07.

[9] V. Ram, A. Das, et. al, "Text-dependent speaker-recognition using one-pass dynamic programming", *Proc. ICASSP'06*.

[10] 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.

[11] A. Das et. al, "Usefulness of Text-Conditioning and A New Database for Text-Dependent Speaker Recognition Research", Proc Interspeech-2008.