interspeech 2008 incorporating SST 08

Usefulness of Text-Conditioning and A New Database for Text-Dependent Speaker Recognition Research

Amitava Das [1], *Gokul Chittaranjan*[2] *and Gopala K. Anumanchipalli*[2]

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

Abstract

Text Dependent (TD) Speaker Recognition systems assume that the password to be uttered by the speaker is known to the system. As the password is known, the system can apply a password-specific model capturing the speaker dynamics well. This enables TD systems to perform better than text-independent systems. We present a variation of the TD systems, called text-conditioning, in which the password is uniquely chosen by each user. This delivers a higher level of discrimination since the linguistic and phonetic differences of the passwords themselves are exploited in separating the speakers. As the database for such a study was not publicly available, we built an extensive database for speaker recognition having such text-conditioning property. The database is tested with various speaker recognition trials. The results indicate that for the design of a practical TD speaker-recognition system, "text-conditioning" does offer a significant edge.

Index Terms: Text-Dependent Speaker Recognition; Speaker Recognition Database; Text-conditioning;

1. Introduction

The two main approaches of automated speaker recognition 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 as a bag of symbols. Therefore, speaker models in the TI methods are distributions in the feature space, modeled by VQ codebooks [12,13] or by Gaussian mixture models [9] built from the extracted features from training speech data. During testing, the TI speaker recognition system tries to find which speaker model (distribution) the test feature-vector-set came from. Such distributions are often overlapping, especially if the password phrases are same or similar for all speakers, leading to lower performances of the TI systems.

Text-dependent (TD) speaker recognition methods [10,11] 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 models can be the stored templates of feature vector sequence as used in the TD method using DTW [11] or they can be HMMs trained by a large number of passwords uttered by the speaker [10]. 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. This important aspect of speaker identity is captured by TD systems and therefore TD systems typically offer much higher performance than the TI systems.

In this paper, we emphasize a variation of the text-dependent speaker recognition approach, we call "text-conditioning", in which essentially the passwords are uniquely selected by each user. The concept of password or a unique stream of words or digits are quite common, accepted and welcome concept to all of us who are using various access control systems in our daily life. The passwords can be multilingual as well. There can also be multiple passwords, which can be "non-wallet" information (information which one typically not carry in his wallet and which are easily remembered) as well as answers to simple questions about things which the user my only know. In such a text-conditioned speaker recognition system, there is an additional edge: The linguistic and phonetic differences of the passwords themselves offer an additional discrimination in the speaker separation process. As a result, there can be two situations for an imposter trial: a) "unknown-password" imposter trial in which the imposter have no idea about the client password and trying something random while trying to get into the system and b) "known-password" imposter trial where the imposter may have overheard the password of the client and uttering the password of the client. Our results will show that for the "unknown-password" situation, even a simple technique such as VQ used typically in TI systems, can offer near-zero or zero error when such text-conditioning used. For the tougher known-password case, such simple VQ methods using text-conditioning (we will call these method TCVQ or text-conditioned VQ) offers much better results overall than the conventional text-independent VQ (to be called here as the TIVQ method). There is no difference in any design or system parameter, just the text-conditioned database itself offers this advantage by which TCVQ outperforms TIVQ, offering performance closer to the more complex dynamic classifiers such as DTW or HMM.

Our objective was to study how such text-conditioning impacts the performance, whether it helps in increasing robustness against imposter attack, whether it simplifies classification scheme. However we could not find any publicly available database which can provide: a) a large speaker-population in which each speaker is speaking a unique password and several versions of recording of the unique password for each speaker exist, b) imposters saying the passwords of client speakers (password-known) as well saying other passwords (password unknown case) c) multiple multilingual passwords per speaker, and d) multiple recording sessions of above. Therefore, for past 15 months we have building such a text-conditioned speaker recognition database at MSR-India (to be referred here as the MSRI database). In this paper we present the details of this database, which we

intend to offer to the global research community and plan to make publicly available for research purpose.

Our paper is organized as follows: Section 2 presents a brief overview of the various database that support conventional TD speaker recognition experiments. Section 3 presents the details of the MSRI database. Section 4 presents the concept of text-conditioning and TCVQ and other speaker recognition experimental set up and details of various methods we applied to show the impact of text-conditioning. Section 5 presents the results, conclusions and our future plans for this study and database collection.

2. Overview of Current Text-Dependent Speaker Recognition Database

Table 1 presents a number of contemporary speaker recognition databases details of which can be found in [1-7]. Note that only a selected few supports text-dependent speaker recognition.

DB Name	TI/TD	Lang.	Password	#Speakers
YOHO	TD	English	6-digit numbers	156 male, 30 female
CSLU V1.1	TI	English	Isolated word	91 speakers
LILA HindiL1	TI	Hindi	Isolated word	2000 speakers
POLYCOST	TD & TI	English	7-digit, sent. & paragraph	74 Male, 60 Female
SIVA	TD & TI	Italian	Digits, personal information	18 male 16, female; 600 imposters
Gandalf	TD & TI	Swedish	sentences, 7digit no	96(48 male, 38 female)
AHUMADA	TD	Castilian Spanish	Isolated digits, 10digit No.	104 speakers

Table 1: Various Speaker Recognition Database

As mentioned earlier, for practical speaker recognition system based on text-conditioning, we need unique passwords per speakers. The passwords from speaker to speaker should differ as much as possible. To build and test such a system we need several versions of the same password for each user and some utterances to simulate known and unknown password imposter attacks. As we could not find a public database which can support the above conditions, we decided to build our own MSRI database presented next.

3. The MSRI Speaker Recognition Database

The MSRI database has **344** speakers, recorded in an office environment over a period of 9 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),

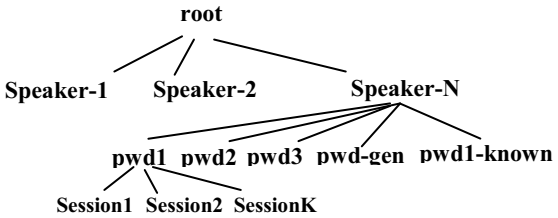
b) PWD-2: 4 words pass-phrase in Indian languages (mother tongue of the speaker)

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.

Not all of the 3 type of passwords are recorded by each user. Each user also said the passwords of a number of other users as well as some random speech material including random 4-digit passwords which are not used by any other users. Here are the statistics:

PWD-1 only: 344 users (277 males & 67 females)
PWD-1 & PWD-2: 94 users
PWD1 & PWD 3: 158 users
All three PWD's: 88 users
Total number of files: 18012
Total amount of recorded speech: 687 minutes
Average amount of speech per user: 2 minutes

All recordings were done in an office environment using PC and regular headsets. The database therefore has realistic office background conditions with SNRs ranging from 2 to 85 dB. The database is laid out in the manner shown in Figure 1.



pwd1/2/3 *sub-directories have various versions of the unique passwords of type 1/2/3 uttered by speaker-N;*

pwd-gen *is a subdirectory which has various utterances of speaker-N which are not any password but random text or digits;*

pwd1-known *is a sub-directory which contains versions of other speaker's (imposter) utterance of the pwd1 of speaker-N*

Figure 1: File Layout of the MSRI database

This allows for the following speaker-recognition trials:

1) Speaker Identification with single or multiple passwords
2) Speaker Verification with single or multiple passwords:
   a) Target trial: test files taken from pwd1/2/3
   b) Unknown-password imposter trial: test files from other speakers' pwd1/2/3/gen files
   c) Known-password imposter trial: test files from speaker-N's pwd1-known files

Therefore, the MSRI database allows for text-independent as well as text-dependent trials and allows simulation of multiple password speaker recognition trials. Most importantly the MSRI database allows the simulation of imposter attacks where the imposter may have overheard the client's password (known-password case) as well as the unknown-password case when the imposter does not know the client's password and thereby making a random guess. As mentioned earlier, the MSRI database is specially designed to study the impact of "text-conditioning" i.e. the impact of using unique password per user.

The MSRI database also allows the study of speech production, accents and style variations as there are many instances of same text being spoken by several speakers. It is also useful for voice transformation research.

4. Text-Conditioned Vector Quantization and Other Speaker Recognition Methods

In this section we illustrate the performance of the conventional text-dependent speaker recognition methods such as multi-template DTW[11] and HMM[10] applied to the MSRI database. Later we present the text-conditioned VQ approach and compare its performance with the rest. In all cases, except VQ methods, 39 dimension MFCC+Delta's were used as feature. For the VQ methods only direct MFCCs were used, their dimensions ranging from 8-12.

Multi-template DTW: Here T=1 to T=4 training templates were used. Optimal choice of global and local constraints was used. Details can be found in [11].

Password-HMM-based Method: Here T=1 to T=6 password templates were used for training the HMM model for the speaker. The model assumes a left-right topology with no skip states and the output distribution at each state is captured as a mixture of Gaussians. The number of states (S) and the number of Gaussians (M) were varied to seek optimal performance. Note that this approach can be considered as a whole password or a sentence HMM as opposed to the HMM model in [10] which was based on speaker-specific models for constituent digits. For training the HMM, the training samples of the passwords are automatically segmented using segmental K-means. An iterative k-means algorithm is employed within each segment to model the distribution as a mixture of Gaussians. Based on the number of instances provided for training, an optimal configuration of the HMM (number of states, number of Gaussian components per state) is arrived at by observing the SID and SV performances on a development set. Table 2 presents the performance of the various configurations of the DTW and the HMM methods.

DTW	SID %err	SV %ERR KnownPwd	HMM	SID %err	SV %ERR knownPwd
T=1	4.78	8.16	T=1;S=4;M=4	20.4	7.3
T=4	0.9	4.89	T=6;S=12;M=4	0.3	2.3

Table2: Performance of the DTW & HMM TD methods

As expected the both methods work better with more training/ templates. The HMM method outperforms DTW method.

Text-Conditioned VQ method: VQ-based speaker recognition methods [12,13] are normally used in text-independent speaker recognition method (we will refer to this as TIVQ) where the speaker-specific VQ codebooks are generated from the speech material of each speaker uttering any random text material. The phone-content of one speaker's training material can be similar to that of another speaker. This creates overlapping distribution of speakers in the feature space, as a result of which TIVQ method offers limited performance in speaker recognition trials. TIVQ also needs a large amount of training material to capture all the phone content and speaker-characteristics. TIVQ requires large codebooks (greater than 256 code vectors per codebook) and a long test sequence [more than 10 seconds] to deliver decent performance.

The text-conditioning effect in the MSRI database created by the usage of unique and distinctly different passwords be each user, allows the same VQ method to deliver better results. All we are doing here is using these unique passwords to train the user-specific codebook [as opposed to random speech material in TIVQ]. There is no other difference other than the training material. We call this unique password-based VQ approach text-conditioned VQ or TCVQ.

The training of codebooks by a text-conditioned training material (several version of the same password) gives the following edges to TCVQ: a) the acoustic feature space of each user become more separated from each other and b) the speaker-specific clusters becomes narrower or more confined in the feature space, c) this enables the TCVQ codebooks to deliver good performance with a significantly fewer code-vectors than TIVQ.

Figure 2 illustrates this with an example MFCC component clusters, showing the scatter-plots of two MFCC components drawn from two password utterances of two speakers. In the left plot, they are drawn from two versions of two random passwords (conventional TIVQ case) and in the right plot they are drawn from the two versions of unique passwords uttered by two speakers (TCVQ or text-conditioned case). Note, that the text-conditioning is confining and separating the clusters.

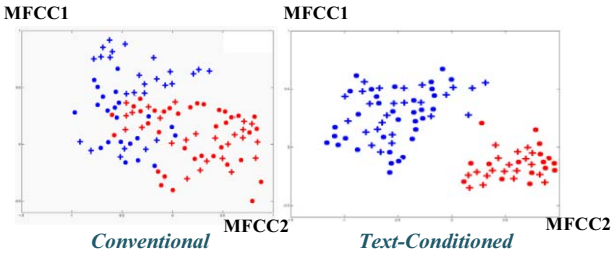


Figure2: Scatter plots of 2 MFCC components; Blue dot/cross → two utterances of speaker-1, Red dot/cross → two utterances of speaker-2. Note that text-conditioned data are separable

Thus, TCVQ offers much higher performance with smaller codebooks than conventional TIVQ and requires lesser amount of training data as shown in Table 3 and 4. A 8x12 TCVQ offers 99.5% SID accuracy, while TIVQ struggles even with 64x12 size codebook offering only 69% accuracy. Even using a single password for training TCVQ offers 90% accuracy with 8x12 size codebook, while even with 6 training passwords TIVQ offers only 70% accuracy with 32x8 size codebook.

CB Size	TIVQ			TCVQ		
	K=8	K=10	K=12	K=8	K=10	K=12
N=4	49.2	43.6	45.1	97.7	98	98.3
N=8	50.7	55.8	58	99	99.3	99.4
N=16	59.1	64.4	65.9	99.1	99.5	99.5
N=32	62.4	68.8	69	99.1	99.3	99.4
N=64	64.1	68.5	69.2	98.9	99.2	99.4

Table3: TIVQ and TCVQ SID performance: Impact of codebook size

Train Size	TIVQ (NxK)			TCVQ (NxK)		
	8x12	16x8	32x8	8x12	16x8	32x8
TR=6	64.9	66.1	70.24	99.5	99.4	99.5
TR=4	58	59.1	62.4	99.4	99.1	99.1
TR=2	42.47	41.8		98.2	97.7	
TR=1	27.2			89.9		

Table 4: TIVQ and TCVQ SID performance: Impact of Training

Finally, we compare all 4 methods and their performance and complexity of operations and memory requirements are shown in Table 5.

		DTW	HMM	TCVQ
Performance Comparison	SID results in % Error	0.9	0.3	0.6
	SV results in % EER For known PWD	4.89	2.3	2.1
Complexity Comparison	Storage: numbers to store per speaker	15.6k	3864	96
	No. of MPY-ADD per trial	12.5M	1.17M	240

Table 5: Performance Comparison of all Four Speaker Recognition Methods using the MSRI database

***Notes**:*

1. Trial Details: 344 speakers; 4744 speaker-identification trials; 5591 speaker verification trials – 847 known-password imposter trials and 4744 unknown-password imposter trials; only PWD1 is used here;

2. Feature used: 12 dimension MFCC for TCVQ; 39 dimensions MFCC+Delta for all other methods;

3. Training Size: 4 passwords are used for training for all except HMM.

4. DTW method: Proper optimization is done by choosing appropriate global constraint; simple local constraint is used; 4 templates used; Details in [11].

5. HMM method: 12 states; 4 mixtures per state; 6 passwords used for training

6. TCVQ: Codebook size 8x12; See [12] for details; Here training is done with pwd1 (4 files), i.e. text-conditioning is applied.

As seen in Table 5, text-conditioning does give an edge in the speaker recognition process as all methods are giving good performance while using a small training set (on average only about 4-6 seconds of training material) and 1-2 second of test material.

The interesting result to note is that with text conditioning, a simple VQ based method, the TCVQ, is delivering performance competitive to the conventional TD methods like DTW and HMM. However, the complexity of TCVQ is much lower than that of either DTW or HMM.

5. Conclusions and Future Directions

We presented a new database exclusively designed for text-dependent speaker recognition and primarily to study the impact of what we call text-conditioning, namely the use of unique password per user. Using unique password is a well-accepted norm among users of all access-control systems, however the use of unique, preferably multi-lingual, phonetically rich and discriminative password, allows the speaker recognition system to be more robust. To our knowledge, this is the only database which provides a large user population in which each user is uttering a set of unique passwords and there are also imposters uttering other's passwords as well as random text. This makes this database an important aid for speaker recognition research, especially to study the impact of text-conditioning or the usage of unique password per user.

Such text-conditioning allows us to exploit the natural phonetic and linguistic difference of the passwords themselves to give an additional edge to speaker discrimination. Text-conditioning not only helps conventional text-dependent method to perform better but as shown here it elevates the performance of this lower-complexity TCVQ method to the level of well-known high-performance and high-complexity conventional TD methods such as DTW and HMM.

The MSRI speaker recognition database allows one to explore the impact of text-conditioning on various other speaker recognition methods as well. For example, a companion paper submitted to this conference by us introduced a novel compressed time-frequency representations of passwords [14] and demonstrated the utility of this new method using this database. As the MSRI database has multiple passwords per user, it also allows researchers to explore newer more robust speaker recognition systems using multiple passwords. The MSRI database is being expanded as an ongoing project. We are expanding the multi-session part of the database as well as the known-password imposter section of the database.

6. References

[1] John Godfrey, David Graff, Alvin Martin - Public Databases for Speaker Recognition & Verification, ESCA workshop on automatic speaker recognition, identification and verification.

[2] Ronald Cole et. al , "The CSLU Speaker Recognition Corpus", in Proc, ICSLP-1998

[3] Pewvska, D., et al. "Polycost: A Telephone Speech Database for Speaker Recognition", Speech communication, 31 (2-3) 2000.

[4] Falcone et. al., "The "SIVA" Sspeech database for speaker verification: Description and evaluation. Cost250 meeting, 1996.

[5] Håkan Melin, "GANDALF - A Sweedish Telephone Speaker Verification Database, Proc. ICSLP 96, pp1954-1957

[6] Ortega-Garcia et. al, "AHUMADA: a large speech corpus in Spanish for speaker identification and verification", ICASSP-98.

[7] Melin H. , "Databases for Speaker Recognition: Activities in COST250 Working Group 2", Proc. COST250 Workshop on Speaker Recognition in Telephony, 1999.

[8] Campbell, "Testing with the YOHO CD-ROM Voice Verification Corpus", in Proc. ICASSP-95, pp341-344.

[9] D. Reynolds, "an Overview of Automatic Speaker Recognition Technology", Proc ICASSP-2000. Pp 300-304.

[10] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," Proc. ICASSP, pp. II-391-394, (1993)

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

[12] F.K.Soong, A.E. Rosenberg, et al, "A vector quantization approach to speaker recognition", AT&T Tech. Journal, Vol 66, pp 14-26 (1987)

[13] A. Das & P. Ghosh, "Audio-Visual Biometric Recognition by Vector Quantization", IEEE SLT-06, 2006.

[14] Amitava Das & Gokul Chittaranjan, "Text-Dependent Speaker Recognition by Efficient Capture of Speaker Dynamics in Compressed Time-Frequency Representations of Speech", submitted to Inter-speech 2008.