MULTILINGUAL SPOKEN-PASSWORD BASED USER AUTHENTICATION IN EMERGING ECONOMIES USING CELLULAR PHONE NETWORKS

Amitava Das, Ohil K. Manyam[1], Makarand Tapaswi[1] and Veeresh Taranalli[1]

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

ABSTRACT

Mobile phones are playing an important role in changing the socio-economic landscapes of emerging economies like India. A proper voice-based user authentication will help in many new mobile based applications including mobile-commerce and banking. We present our exploration and evaluation of an experimental set-up for user authentication in remote Indian villages using mobile phones and user-selected multilingual spoken passwords. We also present an effective speaker recognition method using a set of novel features called Compressed Feature Dynamics (CFD) which capture the speaker-identity effectively from the speech dynamics contained in the spoken passwords. Early trials demonstrate the effectiveness of the proposed method in handling noisy cell-phone speech. Compared to conventional text-dependent speaker recognition methods, the proposed CFD method delivers competitive performance while significantly reducing storage and computational complexity – an advantage highly beneficial for cell-phone based deployment of such user authentication systems.

1. INTRODUCTION

In emerging economies and developing countries proper user-authentication is a highly important component for effective operation of various financial services as well as governmental operations, especially since accountability is low and corruption is rampant. Often money or other forms of aid do not reach the person in need. Most of the time there is no proof that such disbursements have been received. These problems are uncommon in the developed world where various services exist which can effectively authenticate the user using phones, ATMs and internet-ready computers. In emerging economies, proper user-authentication can impact the socio-economic landscape tremendously, but we also need to understand the availability of infrastructure as well as social norms before we choose a method suitable for user-authentication.

In this paper, we present our exploration and evaluation of a cell-phone based user-authentication system in India using a novel speaker recognition method which demonstrates effectiveness in delivering reasonably good performance while facing the tough challenge of processing noisy cell-phone speech. We chose cell-phone as a platform for user-authentication mainly because of the rapid deployment and acceptance of cell-phones amongst common people in India.

India is an amazing country with one foot in the 21st century and another in the middle ages! There are significant strides in scientific and economic development yet a vast percentage of the total population is illiterate and under the poverty line. A majority of these economically challenged people in India became cell-phone savvy in the last 5 years and discovered ingenuous ways to use their cell-phones to enrich their earnings. Fishermen are finding which port to come back to get the maximum price for their fresh catch. Farmers are finding ways out of the traps of middlemen by finding better prices in nearby cities. There are simply too many examples to quote. At present India has about 50 million PC and internet users but there is a large 500 million plus mobile-phone user base which is growing at the rate of 30-40% each year.

MSR-India has a large research team focusing on emerging markets, exploring various socio-economic impacts of technology. Part of our research includes the study of effective user interactions and cell-phone-usage patterns of illiterate and poor people. Researchers here are also exploring mobile phone based commerce, specifically new banking applications using cell-phones which are becoming popular in countries such as Malaysia and South Africa. The mobile-phone based user-authentication system proposed in this paper fits nicely to these research activities at MSR-India as well as other research being done by the growing SLT-D (spoken language technologies for development) community [1-4]. A proper user-authentication system will certainly help many of these SLT-D applications, and in India, mobile phone seems to be the best platform.

This motivated our research and in this paper we present some of the early results of our newly-proposed speaker recognition method which is found to be quite effective for such mobile deployment. Our paper is organized as follows: The basic approach is presented in section 2. Details of our field trials and the mobile speakerID database creation are presented in section 3. Section 4 presents our FGRAM-CFD based speaker recognition method, the performance of which is presented in Section 5. Finally section 6 presents the summary and conclusions.

978-1-4244-3472-5/08/$25.00 ©2008 IEEE

2. SPEAKER RECOGNITION WITH MULTI-LINGUAL PASSWORD USING CELLULAR PHONE

For the mobile phone based user-authentication, we are considering two modes of operation: a) Terminal-side authentication and b) Server-side authentication.

In the first mode, both user-enrollment and authentication is done on the mobile phone. User-specific data (the biometric signatures from the spoken password) is kept on the phone and the enrollment as well as authentication programs are embedded in the cell-phone memory and run on the processors of the cell-phone. This requires the user-authentication method to be of extremely low complexity in terms of storage and computation.

The second approach collects the spoken passwords as voice-data and either extracts relevant features and sends the features as data; or, in places where data services do not exist, sends the voice data to the server where an authentication software does both enrollment and authentication. This approach removes the burden of processing from the mobile phones and transfers the task to the server. It can scale up to as many users as the server can handle. But it suffers from two problems: a) the mobile phone service should be up and running b) the voice data is impacted not only by environmental noise but also by poor network condition which often severely impacts voice quality, especially with burst-type errors. The first approach has the innate advantage of using a mobile phone as a cheap alternative to a PC but it requires sophisticated embedded-programming of the application on the phone itself, requiring intervention from the manufacturer. The second approach can easily be created as a new application-software and can be launched by the service provider. We adopted the second approach for our initial trial.

India is a multi-lingual country and we exploited this in designing our system in several ways. We allowed users to select their own password, having up to six pass-phrases, in their own languages. This created passwords of various phonetic content and the difference between one password to another is exploited in our method. For example, the number “19” is spoken as: ***“oo-neesh”*** in Bengali, ***“hattombattu***” in Kannada, and ***“pantommidi”*** in Telugu. The number “47” is spoken as: ***“saat-chol-lish”*** in Bengali, ***“nalavat-yelu”*** in Kannada and ***nalabhaiyedu”*** in Telugu. Thus even if two users from different parts of India pick “19-47-98” as a password, the spoken versions of the password will be different. The interface also became easy-to-use as one can use his mother tongue. Even if the spoken password is heard, unless the imposter knows the specific language -it will be difficult to pronounce it properly.

One main problem of voice-based user authentication is that others can hear what the user is saying. Therefore, there are concerns about the robustness of the system. We studied the performance of the proposed system with 3 metrics: a) Target case (% accuracy), where we measured what percentage of times a true user is able to get in, b) Imposter case (% Equal Error Rate or EER%) for an imposter who does not know the target-user’s password and is trying something at random, and c) imposter2 case (EER%) -- for an imposter who has heard the target-user’s password, and tries to mimic it as the target-user.

Finally, we have adopted a multiple-phrase password scheme. A longer password having, say, six pass-phrases per password, has the advantages that even if someone overhears you, it is difficult to remember all six pass-phrases. The proposed MSRI method benefits from multiple pass-phrases and the performance improves as more pass-phrases are used. But for the user, it is also quite taxing to remember six words. However, one can design authentication systems of various performance levels using different number of pass-phrases depending on the performance need. For example, for transactions involving smaller amounts of money, a more user-friendly but less imposter-proof system can be created using fewer pass-phrases. For higher amounts, larger number of pass-phrases will provide higher performance and robustness. We describe our mobile speakerID database next.

3. DETAILS OF FIELD TRIALS AND MOBILE-PHONE BASED SID DATABASE COLLECTION

Our project has two phases: a)1st phase: data collection and finalization of the authentication method, b)2nd phase: field trial with terminal-side and server-side adaptations. At present we are creating a mobile-phone speakerID database with appropriate voice samples from mobile-phone users from all over India. We plan to cover at least 100 users per language and plan to cover all 26 official languages of India. This is a huge effort and will span over months of data collection in collaboration with several Indian universities and research organizations.

We are using the most commonly used mobile phones in India for the data collection. Each user is given a phone number to call and a simple “recording-sheet” to fill up in his/her mother tongue following some instructions. The text contains the following passwords a) universal password (to explore text-independent methods), b) unique password selected by users, c) random password (imposter case) and d) passwords of other users (imposter2 case). We built a Dialogic telephony card based IVR system at MSR-India which interacts with the callers and collects the raw speech data which is then processed to create the MSRI mobile-SID database. Each of the various types of passwords is uttered several times. No more than 30 passwords in total are recorded in a single session. We are also collecting data for multiple sessions out of a fixed number of users. We plan to make this database publicly available for research purpose. At the time of writing this paper, we have an early version of our MSRI mobile-SID database which contains Bengali (78 speakers recorded/51 processed), Kannada(24/17), Hindi(20/7) and Telugu(10/4) or a total of 132 recorded speakers out of which 79 are processed at present and used in the evaluation trials presented here.

4. PROPOSED SPEAKER RECOGNITION METHOD

Prevalent speaker recognition methods [5,6,8] extract a sequence of feature vectors (such as MFCC) from the spoken password and then handle them as a feature-sequence or a bag of features. One significant novelty of our proposed method is to create a visual representation of speech or treat speech as a set of images, which we call "Featurogram" or FGRAM. These FGRAMs capture the speaker-identity or the speaking style of a person efficiently by capturing speaker-specific speech dynamics typically exhibited in the co-articulation of various sound units. Given a speech segment of N_1 frames, the N_1 x L FGRAM is formed by extracting an L-dimensional feature vector from each frame and then stacking them up. Thus FGRAM essentially captures the time-varying dynamics of the particular feature. Figure 1 shows an example for FGRAM formation using MFCC. One well-known speech visualization approach is the use of "spectrogram". The proposed FGRAMs are inspired by spectrograms; the novelty of the method proposed here is the automated use of these "images" for speaker recognition.

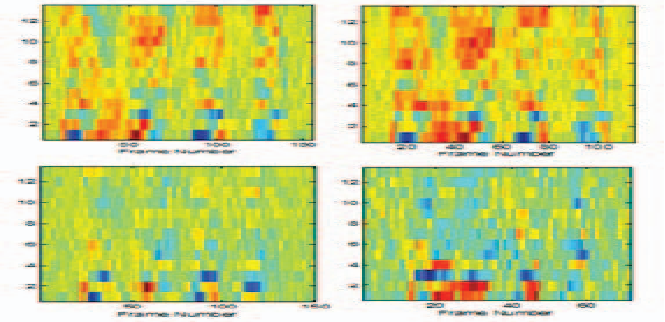


Figure 1: MFCCgram of two utterances of the password by client (top row) and the same password spoken by two imposters (bottom row)

The FGRAMs are next processed by two more steps: a) Sinusoidal model based Time-normalization [7] (Figure 2) and b) compression by Discrete Cosine Transform (DCT) to form a "compressed feature dynamics" (CFD) vector (Figure 3). Details of the FGRAM-CFD concept and methodologies and the various FGRAMs successfully used for speaker recognition can be found in [9-12].

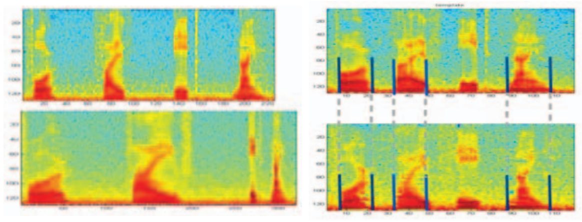
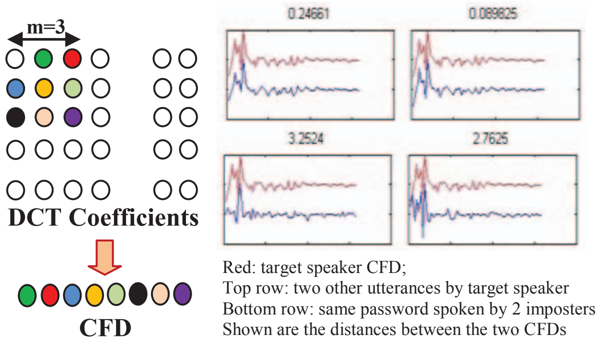


Figure 2: FGRAM normalization; Left: two utterances of the password by client speaker; Right: the same FGRAM time-normalized (see [9-12])

Thus in the proposed MSRI method, the spoken password having up to six pass-phrase is converted to six images or FGRAMs and eventually six CFD signature vectors. MFCC is used to create the FGRAMs.



Red: target speaker CFD;
Top row: two other utterances by target speaker
Bottom row: same password spoken by 2 imposters
Shown are the distances between the two CFDs

Figure 3: CFD formation by DCT and example CFDs. Note the intra-speaker similarity and inter-speaker difference at CFD level

The proposed FGRAM-CFD method is quite convenient for a mobile platform which demands low computational and storage requirements. First of all, the entire password of a speaker, say of 100 frames, can now be represented by only 35 numbers (35 being the CFD dimension) as opposed to 3900 numbers for conventional MFCC+DTW approach [6]. This makes storage requirement significantly lesser for the MSRI approach. Secondly, as different-length passwords are represented in our proposed method by fixed-dimension CFDs, complex dynamic programming based classifications common in conventional DTW or HMM based text-dependent speaker recognition methods, are no longer required. Simpler nearest neighbor classifiers can be used, making the classification part of our method much simpler, reducing complexity significantly (see Table 4).

To further reduce the complexity, a two-stage approach is used. A low-complexity MFCC-VQ based method (see [8] for details) is used as the 1st stage. If the score of the 1st stage is less than a threshold (**TH_L**) then the speaker is accepted, if it is higher than a threshold (**TH_H**) then the speaker is rejected, else it is processed by an FGRAM-CFD based 2nd stage. The scores are computed as follows:

For the first stage, $\boldsymbol{R_1}$ = D1/D2, where D1 is the distance (see [8] for details) of test-feature sequence to the target speaker codebook, and D2 is the distance of the test feature sequence from next-best-speaker codebook.

The score for the 2nd stage $\boldsymbol{R_2}$ is similar, only here distances are computed between the CFD of the test speech and those of the training templates of the target speaker and next-best speaker. See [9-11] for more details.

The final score of the 2nd stage is computed as:

$\boldsymbol{R_{2f}} = R_{21}$ x R_{22} x R_{23} x R_{24} x R_{25} x R_{26} where R_{2k} is the $\boldsymbol{R_2}$ score computed with the FGRAM-CFD of the k-th pass-phrase of the password.

When the second stage is used, this final score **R_{2f}** is compared with a pre-defined threshold to make the accept/reject decision.

5. EVALUATIONS AND RESULTS

The proposed MSRI method is compared with a conventional DTW+MFCC based text-dependent speaker recognition method [6] using an early version of the MSRI mobile -SID database containing 79 speakers, each saying 8 target passwords as well as passwords of 4 other users. Out of these 8, up to 3 are used for training and remaining ones are used for testing. Table 1 compares the performance for DTW baseline with VQ, individual MFCCGRAM-CFD and the combined MSRI method. Table 2 and 3 show the impact of the number of pass-phrases per password and the number of training templates on the performance. Table 4 compares the complexity of MSRI method with DTW.

	Target (%)	Imposter (EER%)	Imposter2 (EER%)
DTW	87.50	0.00	2.68
VQ	96.90	0.22	1.12
MFCCGRAM	99.53	0.00	1.49
Combined(MSRI)	**100.0**	**0.00**	**0.28**

Table 1: Speaker Recognition Performance by various methods[1]

Note[1]: CFD size = 35; no of training templates = 3; no of pass-phrase/password = 6; 1st-stage VQ: codebook size is 16x9; THL=0.8 and TH_H=1.2; Trial no: Target= 395; Imposter = 790; Imposter2 = 655; 39 dimension MFCC is used for DTW.

No. of pass-phrase	1	2	3	4	5	6
DTW	76.58	86.24	88.01	88.48	85.71	87.50
MSRI	89.82	96.02	97.34	97.78	98.23	100

Table 2: Impact of no. of pass-phrase on target-case accuracy (%)

Method	No of templates	Target (%accuracy)	Imposter (EER%)	Imposter2 (EER%)
DTW	T=1	82.14	0.49	3.98
	T=2	87.02	0.38	3.15
	T=3	87.50	0.00	2.68
MSRI	T=1	94.34	0.95	2.25
	T=2	98.90	0.18	0.25
	T=3	100	0.00	0.28

Table 3: Impact of no. of training templates on performance

Method	Computation	Storage
DTW	O(1690650)	O(83130)
MSRI	O(456)	O(2334)

Table 4: Comparison of classification methods in terms of complexity[2]

Note[2]: Computation is measured in terms of multiply-add and storage in terms of number of floating-point numbers to store; Average length of each segment: 85 frames Other assumptions are same as given in Table 1

These results lead to the following observations. All the methods work well for the imposter case showing the merit of our strategy of using multi-lingual unique passwords. Having more pass-phrases and more training templates also boosts the performances. The proposed MSRI method outperforms the baseline DTW for both the target and the imposter2 case (known-password) by a big margin. The proposed MSRI method also offers significantly lesser complexity than the baseline DTW method. These beneficial features make the proposed MSRI method an ideal candidate for deployment on mobile phone platforms.

Note that the evaluation results presented here were run on an early version (79 speakers) of our mobile-SID database. So the results shown here are only indicative of a trend. However, the methods proposed here are already evaluated and reported in [9-11] using a much larger (300+ speaker) MSRI speaker ID database [12] recorded on PC and we have seen similar trends. The FGRAM-CFD method did deliver similar high performance on this 300+ speaker PC database, outperforming DTW at significantly reduced complexity. Therefore, we are quite confident that the proposed MSRI method will work well during the actual field-deployment on the mobile platform and deliver the performance trends shown here.

6. SUMMARY AND CONCLUSIONS

Mobile phones are playing an important role in changing the socio-economic landscapes of emerging economies like India. A proper voice-based user authentication will help in many such new mobile-based applications including mobile-commerce and mobile-banking. We presented an effective text-dependent speaker recognition method using a novel speech feature called CFD which has shown promising results in trials done using an early version of the mobile-speaker-ID database we are collecting for this project. Compared to conventional speaker recognition methods, the proposed CFD method delivers competitive performance while significantly reducing storage and computational complexity – an advantage highly beneficial for cell-phone based deployment of such user authentication systems.

8. REFERENCES

[1] Harsha de Silva et al, "Perceived economic benefits of telecom access at the Bottom of the Pyramid in emerging Asia", In Proc conf. of Intl. Communication Association (ICA), May 2008

[2] Dana Diminescu et. al "Mobile-based money transfer: Weaving together financial and migration fluxes", in Proc ICA, May 2008

[3] Donner, Jonathan and Tellez, Camilo. (in press). "Mobile banking and economic development:Linking adoption, impact, and use", *Asian J. of Comm.*, 18(4).

[4] M. Plauché, at. Al, "Speech Recognition for Illiterate Access to Information and Technology". *Proc.ICTD*, 2006.

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

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

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

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

[9] A Das, et al "Automated Speaker Recognition Using Compressed Temporal-Spectral Dynamics Information of Password Spectrograms", Proc. ISCA ITRW, June 2008.

[10] A Das, et al "Text-dependent speaker recognition by compressed feature-time dynamics derived from sinusoidal representation of speech", Proc. Eusipco-08, Aug 2008.

[11] A Das, et al "Text-Dependent Speaker Recognition by Efficient Capture of Speaker Dynamics in Compressed Time-Frequency Representations of Speech", Proc. Interspeech-08, Sep 2008.

[12] A Das, et, al "Usefulness of Text-Conditioning and A New Database for Text-Dependent Speaker Recognition Research", Proc. Interspeech-08, Sep 2008.