PHONETIC CONSISTENCY IN SPANISH FOR PIN-BASED SPEAKER VERIFICATION SYSTEMS

J. Ortega-Garcia[1*], *J. Gonzalez-Rodriguez*[1], *D. Tapias-Merino*[2]

[1] EUIT Telecomunicación, Universidad Politécnica de Madrid
[2] Telefónica Investigación y Desarrollo
[*]Ctra. Valencia km. 7, 28031 Madrid, Spain
e-mail: jortega@diac.upm.es; http://www.atvs.diac.upm.es

ABSTRACT[(**)]

The use of uttered Personal Identification Numbers (PIN) is a well-suited approach for person identification through voice in real applications. In this paper, speaker verification with short 4-digit strings, in a pragmatic perspective where very few utterances for training are available, is accomplished. The problem here arises due to the small quantity of voice available in short PIN utterances. Furthermore, it has to be taken into account the specificity of Spanish in this task, as digit strings are not uttered in a isolated digit-by-digit basis, but mentally grouped without constraints, and read as whole figures, with varying groups for different utterances of the same PIN. This specific factor induces high dependency on the phonetic contents of the PIN, and complicates considerably the design of text-dependent systems. Considering this, a text-independent GMM speaker verification system, including 'nearest reference speaker' and 'universal background model' score normalization, together with CMN channel compensation, has been evaluated over a specific PIN database, where different training conditions (phonetic dependent/independent) are tested.

1. INTRODUCTION

Personal identification through voice for access control and restricted services management, in both remote and proximity implementations, has become one of the most important issues in state-of-the-art speech technology applications [1]. Therefore, the use of PIN utterances perfectly fits for speaker verification tasks in real developments. In this paper, we are not concerned about long PIN codes [2], because if they are really PINs and not passport or telephone numbers (which are very easy to know), they are extermely difficult ro remember. Then, we will use 4-digit strings as PINs, exactly as in credit cards, in a pragmatic perspective where only very few utterances for training are available. We believe that short numeric codes will meet the requirements of many speaker verification applications.

Talking about short PIN codes means taking into account the peculiarity of Spanish speakers (and surely others worldwide) when pronouncing digit strings, in the sense that these are not uttered in a isolated digit-by-digit basis, but mentally grouped without constraints, and read as whole figures [3]. For instance, digit string "1 2 3 4" can be uttered as "one two three four", "twelve thirty four", "one thousand two hundred and thirty four", "twelve three four", and moreover, different for different utterances from the same speaker. This specific factor induces high dependency on the phonetic content of the PIN, and complicates considerably the design of text-dependent systems.

This specific situation has led us to consider the use of text-independent recognition to face the problem of PIN-based speaker verification. In this sense, we have employed a GMM approach for speaker verification through PIN codes. A task-specific speech database, named TelPIN, has been recorded in order to carry out several verification tests, including telephone handset and channel compensation through a Telefonica real-time version of cepstral mean normalization (CMN) [4]. In order to test the system, two universal background models (UBM) have been obtained from specific spanish databases [5], and several score normalization methods, as 'nearest reference speaker' and 'UBM' normalization have also been tested [6].

2. TelPIN DATABASE

TelPIN database has been specifically recorded for this project, containing a total of 50 speakers (25 male + 25 female). Originally, it has been *in situ* recorded through a lapel Sony microphone. For the telephone speech experiments contained in this paper, an artificial high-quality Brüel&Kjær artificial-mouth has been used to reproduce microphone-speech files through real telephone links. This procedure allows to have available exactly the same speech but using different handsets and telephone channels, so channel mismatch can be in each case better studied. Specifically, two Spain-widespread standard handsets have been used, namely TEIDE type (Handset 1), and FORMA type (Handset 2) (these two types cover about 75% of all handsets deployed in Spain). Experiments are designed on the basis of realistic training, where no more than 2 or 3 repetitions can be requested to avoid the speaker from feeling annoyed.

2.1 TelPIN Contents

TelPIN corpus has been specifically designed to contain the training and testing speech material for this task. From now on, we will classify possible pronunciation of an specific PIN as *iso-phonetic* (same phonetic content) utterance of that PIN, and any other pronunciation as *alo-phonetic* (different phonetic content).

Training Speech

- 10 phonetically-balanced (specifically designed and restricted for a PIN application) digit strings of variable

[(**)] This work has been supported by "Telefónica Investigación y Desarrollo" under contract TI+D 12051.

length, common to all speakers. With this material, we can construct PIN pronunciation-independent speaker models.

- 3 repetitions of specific PIN per speaker, all pronounced in an iso-phonetic way, for pronunciation-dependent speaker models with one, two or three PIN repetitions.

Testing speech

- 5 repetitions of iso-phonetic specific PIN per speaker.
- 2 more repetitions of specific PIN per speaker, but this time pronounced in an free alo-phonetic way. In this way, we can test the system with different pronunciations of the correct PIN.
- 2 other PIN, one repetition each, corresponding to other speakers personal PIN code. These utterances will be used to determine how the system operates with deliberate or intentional speakers.

2.2 TelPIN tasks

Training Tasks

- Training task *TR_STR*: 1 GMM per speaker is trained with *10 common* (for all speakers) *phonetically-balanced digit strings*. This means phonetic content is the same for all speakers, and no specific PIN information is used.
- Training tasks *TR_NPIN*: 1 GMM per speaker is trained with *N repetitions of his/her PIN*. As *N* ranges from 1 to 3, three different models per speaker are obtained.

Testing Tasks

- TEST_A: False rejection (FR) curves obtained with 5 repetitions of iso-phonetic PIN utterances. False acceptance (FA) curves obtained with 2 PIN corresponding to *intentional* impostors (same PIN as target speaker).
- TEST_B: FR curves obtained with 5 repetitions of iso-phonetic PIN utterances. FA obtained using all non-target speakers as causal or *unintentional* impostors (pronouncing their own PIN).

3. SPEAKER VERIFICATION SYSTEM

In order to perform speaker verification tests over the available data, a text-independent automatic speaker verification system, based in Gaussian Mixture Model (GMM) approach, has been employed. 8 Mel-frequency cepstral coefficients (MFCC) plus 8 ΔMFCC and 8 ΔΔMFCC, have been used as feature vectors in all cases. Frames of 32 ms. taken every 16 ms., with Hamming windowing and pre-emphasis factor of 0.97 are used as input to the system.

Tests without normalization and with likelihood-domain (score) normalization [1] have been accomplished. As the density at point X (input sequence) for all speakers other than the true speaker, S, is frequently dominated by the density for the nearest reference speaker, nearest reference speaker normalization criterion has been applied:

$$\mathbf{log}L(X) = \mathbf{log}p(X|S = S_C) - \max_{S \in \mathbf{ref}, S \neq S_C} \mathbf{log}p(X|S)$$

where S_c means claimed speaker model. Another normalization criterion has been used, corresponding to the so called universal background model (UBM) normalization, that is:

$$\mathbf{log}L(X) = \mathbf{log}p(X|S_c) - \mathbf{log}p(X|S_U)$$

where S_c means claimed speaker model and S_U means universal (generic) model. The design of the universal model, in the sense of selecting appropriate training material, remains the cornerstone of the whole procedure.

System performance will be provided as average Equal Error Rate (EER) [7], for each case, because we would need too much DET curves [8] to be included in this paper to report all the experiments performed. No separate impostor population is used, so non-target speakers are considered as impostors in each case.

4. EXPERIMENTAL RESULTS

GMMs are statistical models in which underlying temporal structure of speech is lost. This property makes them suitable for the text-independent speaker verification problem. Anyway, if training and testing is accomplished using the same phonetically-specific (short) utterance (PIN, password, passport code, etc), the model will be speaker *and* phonetically specific, hence tending to reject other non-specific utterances. A number of experiments have been performed, and results are presented subsequently.

• Experiment 1: Microphone speech

Experiment 1 establishes benchmark results, as same session, same microphone speech is used in all cases. As stated previously, there are two different training procedures, generating 4 different models per speaker, namely: TR_STR, models trained with 10 common (unspecific) digit strings, and TR_*N*PIN, models trained with *N* repetitions of PIN, *N* varying from 1 to 3. Tests are accomplished considering male and female separate populations, and also using all of them together. No kind of score normalization is used in this test. Table 1 shows results of Experiment 1.

EER(%)	Gender	TR_STR	TR_1PIN	TR_2PIN	TR_3PIN
TEST_A	Male	**4.75**	**2.56**	**0.01**	**0.00**
	Female	**1.14**	**0.14**	**0.00**	**0.00**
	All	**2.70**	**0.22**	**0.00**	**0.00**
TEST_B	Male	**4.23**	**0.57**	**0.00**	**0.00**
	Female	**0.93**	**0.86**	**0.27**	**0.00**
	All	**3.42**	**1.32**	**0.38**	**0.39**

Table 1: Benchmark microphone speech results.

From Table 1, some conclusions can be derived: specific PIN training (TR_*N*PIN) is more effective that generic digit string training (TR_STR). With only 1 repetition of PIN code, results are good enough. With 2 or 3 repetitions, less than 0.4% error is observed. TEST_A (deliberate impostors) and TEST_B (undeliberate impostors) show similar results.

- **Experiments 2 and 3: Telephone speech, matched conditions**

In Experiment 2, Handset 1 ('TEIDE' type) is used for training and testing. Same tests as in Experiment 1. Nevertheless, in this case nearest reference speaker normalization is also shown. Results of Experiment 2 can be found in Table 2.

EER(%)	Gender	TR_STR	TR_1PIN	TR_2PIN	TR_3PIN
TEST_A	Male	**4.3 / 0.0**	**2.3 / 0.0**	**0.0 / 0.0**	**0.0 / 0.0**
	Female	**3.2 / 2.6**	**1.8 / 0.0**	**1.0 / 0.0**	**0.4 / 0.0**
	All	**3.7 / 0.2**	**2.0 / 0.2**	**0.5 / 0.0**	**0.2 / 0.0**
TEST_B	Male	**19.4 / 0.0**	**5.1 / 0.1**	**0.5 / 0.0**	**0.3 / 0.0**
	Female	**18.3 / 0.1**	**2.7 / 0.7**	**0.7 / 0.4**	**0.2 / 0.4**
	All	**12.2 / 0.1**	**3.2 / 0.1**	**0.4 / 0.0**	**0.2 / 0.0**

Table 2: Telephone speech, Handset 1, matching conditions. EERs expressed without/with nearest reference speaker normalization.

In Experiment 3, Handset 2 ('FORMA' type) is used for training and testing. Same tests as in Experiment 2, including nearest reference speaker normalization as in previous table. Table 3 shows results of Experiment 3.

EER(%)	Gender	TR_STR	TR_1PIN	TR_2PIN	TR_3PIN
TEST_A	Male	**2.4 / 0.2**	**0.8 / 0.0**	**0.1 / 0.0**	**0.0 / 0.0**
	Female	**3.3 / 4.6**	**2.6 / 0.0**	**1.1 / 0.0**	**0.5 / 0.0**
	All	**2.9 / 0.3**	**1.7 / 0.0**	**0.6 / 0.0**	**0.3 / 0.0**
TEST_B	Male	**19.6 / 0.4**	**3.6 / 0.1**	**0.8 / 0.0**	**0.1 / 0.0**
	Female	**20.5 / 0.2**	**3.1 / 0.0**	**0.7 / 0.0**	**0.1 / 0.0**
	All	**12.9 / 0.2**	**2.7 / 0.1**	**0.5 / 0.0**	**0.1 / 0.0**

Table 3: Telephone speech, Handset 2, matching conditions. EERs expressed without/with nearest reference speaker normalization

Results found in Experiments 2 and 3, with telephone speech and matching conditions, show how specific PIN training (TR_*N*PIN) is much more effective that generic digit string training (TR_STR). It is also remarkable the fact that testing identical PINs (TEST_A, deliberate impostors) with generic training concentrates only on speaker identity, whereas different PINs (TEST_B, casual impostors) with generic digit string training relies not only on identity but also on phonetic similarity of testing contents, which always occurs with respect to a non-specific phonetic model (TR_STR). It is also shown how 2 PIN repetitions may suffice for obtaining good speaker verification results. The use of score normalization has improved results outstandingly.

- **Experiment 4: Telephone speech, handset mismatch**

In Experiment 4, telephone speech is used, and the effect of training with one handset and testing with other different handset is presented. Table 4 presents results in the form CASE1/CASE2, where CASE1 means training with Handset 1 and testing with Handset 2, whereas CASE2 means the opposite. Nearest reference speaker normalization is used in all cases.

EER(%)	Gender	TR_STR	TR_1PIN	TR_2PIN	TR_3PIN
TEST_A	Male	**1.5 / 0.8**	**0.7 / 0.8**	**0.0 / 0.0**	**0.4 / 0.0**
	Female	**1.3 / 8.1**	**0.3 / 1.7**	**0.0 / 1.4**	**0.0 / 0.7**
	All	**0.4 / 1.5**	**0.4 / 0.6**	**0.0 / 0.7**	**0.2 / 0.4**
TEST_B	Male	**2.2 / 0.7**	**1.5 / 2.2**	**0.0 / 0.0**	**0.0 / 0.0**
	Female	**6.6 / 3.4**	**0.1 / 0.1**	**0.1 / 0.1**	**0.0 / 0.2**
	All	**3.0 / 1.5**	**1.1 / 1.0**	**0.0 / 0.1**	**0.0 / 0.1**

Table 4: Telephone speech, handset cross-mismatching conditions. Score normalization applied in all cases.

The effect of channel mismatching worsens results with respect to Experiments 2 and 3. Anyway, as it can be derived from Table 4, score normalization techniques are also effective for channel mismatching. It is remarkable that from 2 repetitions of PIN for training (TR_2PIN), EER smaller (or much smaller) than 0.7% are obtained.

- **Exp. 5: Telephone speech, handset mismatch, CMN and UBM Normalization**

In this case, when training is accomplished with only Handset 2 information, we are presenting testing results when matching conditions occur (Handset 2), and when mismatching conditions (Handset 1) verify. For the score normalization stage, nearest reference speaker method is still used, but it is also compared with universal background model normalization.

Two different universal models are used. The first of them has been obtained from GAUDI/AHUMADA speech database [4], which is completely independent from TelPIN speech data. The other model is derived directly from TelPIN data. Table 5 shows Experiment 5 results, comparing all these score normalization methods and using in all cases CMN channel compensation.

EER(%)	TEST_*J*	TRAIN	NoNor	Nearest	Gaudi	TelPin
Match	A	TR_STR	4.7	0.1	0.6	1.3
		TR_1PIN	1.5	0.0	0.1	0.3
		TR_2PIN	0.6	0.0	0.1	0.0
		TR_3PIN	0.4	0.0	0.0	0.0
	B	TR_STR	17.7	0.1	2.8	3.0
		TR_1PIN	3.2	0.1	0.3	0.4
		TR_2PIN	0.8	0.3	0.0	0.1
		TR_3PIN	0.7	0.0	0.0	0.0
Mis-match	A	TR_STR	7.1	0.3	1.6	2.4
		TR_1PIN	3.3	0.1	0.7	0.8
		TR_2PIN	1.0	0.4	0.4	0.4
		TR_3PIN	0.6	0.0	0.0	0.0
	B	TR_STR	18.4	0.2	3.9	4.9
		TR_1PIN	5.4	0.2	1.3	1.2
		TR_2PIN	2.0	0.0	0.1	0.1
		TR_3PIN	0.8	0.0	0.1	0.0

Table 5: Verific. results with channel match and mismatch, making use of CMN and several score normalization methods.

In a general manner, it can be said that, without score normalization, CMN will slightly degrade results in matched conditions, and will remarkably contribute to improve results in mismatched conditions. When CMN is combined with nearest reference speaker normalization method, results will improve in all cases, producing in our case verification error rates under

0.3%. Considering all score normalization techniques proposed, nearest reference speaker technique produces the best results. But, as far as this technique in not very realistic in on-line real systems, universal model normalization is an excellent alternative. TelPIN (specific) universal background model, produces very good results for specific PIN training (TR_*N*PIN). Anyway, using external generic universal model from GAUDI database, produces about the same results that using TelPIN universal model, with the advantage that this data can be used in any other generic situation regarding different verification systems.

• Experiment 6: Telephone speech, handset cross-mismatch, alo-phonetic utterances

Experiment 6 is similar to Experiment 5 regarding training stage, accomplished with Handset 2, score normalization procedures and the use of CMN channel compensation scheme. The difference between Experiment 5 and 6 stands on testing tasks. For Experiment 6, two different testing tasks have been developed, namely:

- TEST_C: False rejection (FR) curves obtained with 2 repetitions of speaker PIN, but pronounced in a *alo-phonetic* way. False acceptance (FA) curves obtained with 2 PIN corresponding to *intentional impostors* (same PIN as target speaker).
- TEST_D: FR curves obtained with 2 *alo-phonetic* utterances of speaker PIN. FA obtained using all non-target speakers as *undeliberate impostors* (pronouncing their own PIN).

EER(%)	TEST_*J*	TRAIN	NoNor	Nearest	Gaudi	TelPin
Match	C	TR_STR	6.0	0.1	0.3	0.1
		TR_1PIN	18.3	7.0	15.5	14.9
		TR_2PIN	18.2	8.4	12.8	13.6
		TR_3PIN	16.6	3.8	12.7	13.0
	D	TR_STR	19.6	0.1	1.8	0.9
		TR_1PIN	22.1	1.5	10.8	9.8
		TR_2PIN	18.6	0.1	6.9	6.3
		TR_3PIN	16.8	0.0	6.1	5.0
Mis-match	C	TR_STR	8.3	0.0	0.9	0.2
		TR_1PIN	21.0	9.1	16.3	11.0
		TR_2PIN	18.4	10.0	17.5	10.8
		TR_3PIN	18.6	6.3	13.1	10.8
	D	TR_STR	20.5	0.1	2.5	1.3
		TR_1PIN	23.9	4.0	13.4	11.1
		TR_2PIN	21.9	0.4	8.8	7.6
		TR_3PIN	19.5	0.1	7.8	5.9

Table 6: Verification results with matched channels and channel mismatch, making use of CMN and several score normalization methods, for testing tasks TEST_C and TEST_D.

This experiment demonstrates the high sensibility of text-independent GMM models to the phonetic content of training speech, specially when short utterances are used, as in the case of PIN based verification. If phonetic content of short test utterances (TEST_C and TEST_D) is different from that in training utterances (TR_*N*PIN), results worsen, showing the lack of phonetic consistency. In this case, a generic training (TR_STR) works better, as being more general it is also more adapted to phonetic variations.

5. SUMMARY

In this work, we have shown that uttered 4-digit PIN information can be used for person identity verification through the speaker voice. However, we have to take care with the specificity of spanish pronunciation of PINs, which is usually different for the same speaker in different utterances of the same PIN. Then, as we have a very small amount of training material, we can force the speaker to pronounce their PINs isophonetically, where isophonetic PIN training has been shown more effective than generic digit string training, or free the speaker from any constraint in his pronunciation, using then the phonetically-balanced digit-strings trained models. In this way, phonetic consistency of PIN utterances determine a kind of “text-dependency” (phonetically-specific) characteristic of text-independent speaker modeling through GMMs. Further work in this project include the evaluation with a multisession telephone-speech PIN database, and the combination of the verification system with the Telefonica state-of-the-art alo-phonetic PIN recognition system in order to know *who* says the *correct* PIN.

6. REFERENCES

[1] J.P. Campbell, "Speaker Recognition: A Tutorial", Proc. of the IEEE, Vol. 85, No. 9, pp. 1437-1462, 1997.

[2] J. Gonzalez-Rodriguez, S. Cruz and J. Ortega-Garcia, “Biometric Identification through Speaker Verification over Telephone Lines”, *Proc. of 33rd IEEE Intnl Conf. on Security Technology*, pp. 238-242, Madrid (Spain), 1999.

[3] J. Ortega-Garcia, S. Cruz-Llanas and J. Gonzalez-Rodriguez, “Facing Severe Channel Variability in Forensic Speaker Verification Conditions”, *Proc. of EuroSpeech’99*, pp. 783-786, Budapest (Hungary), 1999.

[4] J. Gonzalez-Rodriguez and J. Ortega-Garcia, "Robust Speaker Recognition through Acoustic Array Processing and Spectral Normalization", *Proc. of ICASSP'97*, Munich (Germany), 1103-1106, 1997.

[5] J. Ortega-Garcia, J. Gonzalez-Rodriguez et al., “AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification”, *Proc. ICASSP-98*, Seattle (USA), pp. 773-776, 1998.

[6] D.A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification", *Proc. of EuroSpeech'97*, Rhodes (Greece), 1997.

[7] D. Gibbon et al., eds., Assessment of Speaker Verification Systems, in *Handbook of Standards and Resources for Spoken Language Systems*, M. de Gruyter, Berlin, 1997.

[8] A. Martin, "The DET Curve in Assessment of Detection Task Performance", *Proc. EuroSpeech'97*, pp. 1895-1898, Rhodes (Greece), 1997.