U-NORM Likelihood Normalization in PIN-Based Speaker Verification Systems

D. Garcia-Romero, J. Gonzalez-Rodriguez,
J. Fierrez-Aguilar, and J. Ortega-Garcia

Speech and Signal Processing Group (ATVS)
Universidad Politécnica de Madrid (UPM)
dgromero@atvs.diac.upm.es
{jfierrez,jgonzalez,jortega}@diac.upm.es
http://www.atvs.diac.upm.es

Abstract. This paper present a new likelihood normalization technique, entitled U-NORM, for speaker recognition systems based on short utterances. A comparison between this new approach and the widely used Z-NORM is reported and evaluated. Phonetic dependency between the speaker model and the test speech utterances is determined as the main impediment for a good performance of Z-NORM technique. A set of experiments are developed on a specifically acquired PIN-oriented real-users database showing the higher performance of the new technique for PIN based security applications. U-NORM provides a common likelihood scale for all system users allowing speaker independent thresholds that simplify the enrollment process and add robustness to PIN based security applications.

1 Introduction

It is generally known that speaker recognition systems provide one of the most feasible scenarios for remote security applications due to wide deployment of access points (landline, cellular telephone and Internet). This leaves the voice signal as a desirable biometric modality for any system with remote secure authentication needs. Furthermore, this ease of access enables fully-automated remote acquisition of large databases and consequently enough data to develop common benchmark's and obtain statistically significant assessment of the speaker recognition technologies.
As a good example we may consider the NIST [1] yearly text independent speaker recognition evaluation whose baseline test happens to be the most commonly used benchmark for speaker recognition systems. This baseline test provides 2 minutes of speech for speaker modeling and 30 seconds for test segments.

Although NIST yearly evaluations have greatly contributed to the development of new speaker recognition algorithms, some of these technologies need to be tuned or even disregarded when applied in a different framework. A good example of this may be found in short utterances based speaker recognition systems where, generally, the available amount of data for training and testing differs significantly from the NIST case.

J. Kittler and M.S. Nixon (Eds.): AVBPA 2003, LNCS 2688, pp. 208-213, 2003.

Besides the amount of data, a new artifact appears in the PIN framework due to the scarcity of acoustic information and phonetic variability of the training data. Such effect results on a high dependency between the speaker model and the phonetic content of the speaker PIN even though the technology used is text independent (GMM [2]). This artifact may be considered beneficial as long as the PIN is only known by the client and not anyone else, which is the most common scenario.

One of the most useful and common techniques that is greatly affected by the phonetic dependency of the speaker model is the Z-NORM likelihood normalization [2]. This normalization approach provides a common likelihood scale for all system clients by means of normalizing the speaker likelihoods with the a priori estimated mean and variance $\{\mu_{IMP},\sigma_{IMP}\}$ from a generic set of impostors selected in the development phase. The a priori statistics are computed once for each client during the enrollment phase and used to normalize the likelihood of the speaker utterances. This strategy provides an approximately zero mean and a unity variance distribution for the impostor's likelihoods, if the a priori estimates are well-adjusted to the real distribution, and higher likelihoods for the client's utterances (the more the similitude between the test utterance and the training utterances the higher the likelihood). Due to the phonetic dependency in PIN based applications, two different classes of impostors may be considered: real impostors (those who know the client's PIN) and casual impostors (those who have no knowledge of the client's PIN). Since the Z-NORM attempts to normalize likelihoods based on the a priori knowledge of impostor's distribution new considerations must be taken into account for PIN based systems.

The main reason for regarding likelihood normalization of relevance importance is the possibility of establishing speaker independent thresholds, which provides two major benefits, namely simplicity of the enrollment process [3] and reduction of storage space in the recognition system.

Due to the above mentioned advantages and the need for new considerations in PIN based frameworks, this paper reports on a series of comparative experiments on likelihood normalization techniques and proposes a new algorithm, entitled U-NORM, that takes into account the specifics of PIN based speaker recognition systems.

2 System Description

Current speaker recognition systems rely almost exclusively on short-time acoustic information. UBM-MAP-adapted Gaussian Mixtures Models [4] represent the state-of-the-art technique in text independent speaker recognition achieving a very good performance but conditioned to the acoustic environment.

2.1 Baseline System

The baseline system is based on our MAP-GMM system used in the 2002 NIST evaluation [4]. A gender-independent 512 mixtures UBM is trained with approximately one hour (gender balanced) of microphone speech acquired in the same

conditions of the database utterances (section 3.1). Features vectors consist of 19MFCC+19ΔMFCC obtained from a 20 ms Hamming window shifted 10 ms. Target speaker models are trained via MAP adaptation of the UBM with 10 iterations. Channel compensation is performed by means of Cepstral Mean Normalization (CMN) and UBM normalization is applied to the speaker likelihoods.

2.2 U-NORM

As stated above, a common likelihood scale for all speakers is something desirable since speaker independent thresholds have many advantages. PIN based applications add new considerations to the likelihood normalization procedure since two different kinds of impostors are considered.

For a proper use of the Z-NORM technique, the subset of impostors used to calculate the a priori statistics must know the PIN of all system clients since the essence of this algorithm lies in the estimation of real impostor likelihoods. This technique is impracticable in online systems since the a priori subset of impostors have no knowledge of the PIN number of the system clients in the development phase. Due to that, only casual-impostors may be used in the a priori estimation of impostor likelihoods yielding a mismatch between the estimated likelihoods and real impostor distributions. The cause of this mismatch lies in the implicit phonetic dependency between the client model and the real-impostor utterance which yields a higher likelihood for this situation than in the casual-impostor case. Hence, real-impostors will score higher than the estimated impostor distribution increasing the risk of obtaining likelihoods beyond the settled threshold.
In consequence, impostor-based likelihood normalization techniques do not seem to fit into the PIN based applications, since the likelihoods of real-impostors remain unknown in the development phase.

A new approach may be considered by substituting the impostor-based likelihood normalization by a user-based. This technique has been named U-NORM and is performed in two steps:

1. Estimation of the outcomes of the client model q with a subset of the client utterances, calculating the mean and variance of the likelihoods distribution $\{\mu_{IMP},\sigma_{IMP}\}$.
2. During the testing phase, after the baseline system outcomes the raw likelihood $\Lambda(X|q)$, the following normalization is performed:

$$\Lambda_{UNORM}(X \mid q) = \frac{\Lambda(X \mid q) - \mu_q}{\sigma_q}. \quad (1)$$

Therefore, in the enrollement phase (either one session or multi-session) some client utterances will be used for training and other for U-NORM normalization.

Table 1. Database structure

# Sesion	1	2	3	4	5
Clients	10	6	12	0	19
Impostors	11	7	12	0	0

3 Experiments

3.1 Database

A total amount of 47 speakers are involved in this database acquired specifically for PIN based applications assessment. All speaker utterances were collected within a one month period of time using a Plantronics headset USB microphone. Up to 5 different sessions were used to collect the data. In the first session all the speakers where asked to utter 5 repetitions of their PIN, an eligible number of real-impostor trials and again two repetitions of their own PIN. In subsequent sessions only two utterances of their own PIN and two real-impostor utterances were requested. The number of sessions in which each speaker was involved is not constant as was the number of real-impostor trials. It is important to remark that when the speakers performed as an impostor not only was the PIN known but also the way it was uttered by the client. The following chart shows the number of clients and impostors that attended to one, two, three, four or five sessions.

3.2 Results

All the experiments were performed with the 47 system clients. The speaker models were trained with three PIN utterances in two different training conditions, namely mono-session (first session utterances) and multi-session (utterances from different sessions). To assess the system performance, false alarm probabilities were always computed with all the real-impostor utterances, whereas miss detection probabilities were computed in two different conditions, namely mono-session (first session utterances of the client) and multi-session (client utterances from different sessions). Combination of all training-testing conditions yields four different possibilities but only three of them were considered since the multi-session training and mono-session testing condition does not report any interesting information.

Three normalization techniques are compared through the application to the raw likelihoods generated by the baseline system:

- Z-NORM with casual-impostors, also named “a priori” since no knowledge of the client’s PIN is necessary. 51 speakers were used as impostors for all clients.
- Z-NORM with real-impostors, also named “a posteriori” since knowledge of the client’s PIN is necessary. It’s important to remark that this approach is not valid for online systems but the results are computed to remark the necessary distinction between real-impostors and casual-impostors.
- U-NORM with the client utterances not used for training.

Table 2 presents a summary of all the experiments results in terms of equal error rate (EER). Baseline system performance is also showed in order to make noticeable the relative improvement of the likelihood normalization techniques.

Table 2. Experiments results in terms of % of EER with different normalization techniques and training-testing conditions

Train-Test condition	Likelihood normalization			
	None	Z-NORM a priori	Z-NORM a posteriori	U-NORM
Mono-Mono	4 %	18 %	4 %	1.5 %
Mono-Multi	7 %	20 %	5 %	3.5 %
Multi-Multi	5 %	19.5 %	5 %	2 %

Figure 1 depicts DET plots for Z-NORM a posteriori and U-NORM in order to allow a more exhaustive comparison for all possible system operating points.

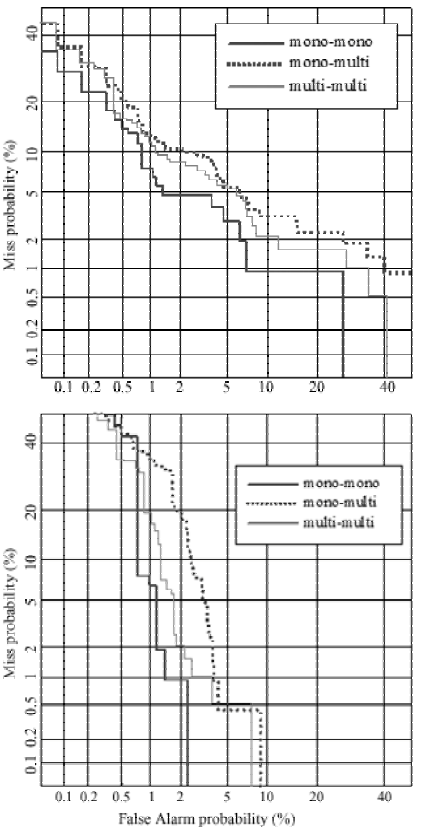

Fig. 1. DET plot for Z-NORM a posteriori *(top)* and U-NORM *(bottom)* techniques in three different training and testing conditions

As shown in table 2, U-NORM technique provides the best results for the PIN based experiments. Z-NORM a posteriori performance is good as well but it is not practicable in online systems. Z-NORM a priori performance is the worst, even worse than raw likelihood results. This is due to the fact that casual-impostors statistics are not representative of the real-impostors likelihoods.

4 Conclusions

This paper reported on a series of comparative experiments on likelihood normalization techniques and proposed a new algorithm, named U-NORM, that takes into account new considerations regarding the PIN based security applications.
Analyzing the results obtained in all the experiments we may conclude that U-NORM technique provides excellent results for PIN based applications allowing the use of a common likelihood scale for all system clients and enabling speaker independent thresholds with a considerable reduction of the enrollment process. Z-NORM technique only performs correctly when used with real-impostor statistics (a posteriori) which are not available for online applications. Casual-impostor statistics are available for online applications but perform poorly due to the phonetic dependency of the PIN based applications.

Acknowledgements

This work has been supported by the Spanish Ministry for Science and Technology under project TIC2000-1669-C04-01. J. F.-A. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research.

References

[1] NIST 2003 Speaker Recognition Evaluation Plan, at http://www.nist.gov/speech/tests/spk/2003.
[2] Douglas A. Reynolds *et al.*, "Speaker Verification using Adapted Gaussian Mixture Models, Digital Signal Processing", vol. 10, pp. 19-41 (2000).
[3] J.-B. Pierrot *et al.*, "A Comparison of a Priori Threshold setting Procedures for Sperker Verification in the CAVE Project", ICASSP'98.
[4] D. García-Romero *et al.*, "ATVS-UPM Results and Presentation at NIST'2002 Speaker Recognition Evaluation", Vienna, VA, 2002.