Weighting Scores to Improve Speaker-Dependent Threshold Estimation in Text-Dependent Speaker Verification

Javier R. Saeta[1] and Javier Hernando[2]

[1] Biometric Technologies, S.L.,
08007 Barcelona, Spain
j.rodriguez@biometco.com
[2] TALP Research Center, Universitat Politècnica de Catalunya (UPC),
08034 Barcelona, Spain
javier@talp.upc.es

Abstract. The difficulty of obtaining data from impostors and the scarcity of data are two factors that have a large influence in the estimation of speaker-dependent thresholds in text-dependent speaker verification. Furthermore, the inclusion of low quality utterances (background noises, distortion...) makes the process even harder. In such cases, the comparison of these utterances against the model can generate non-representative scores that deteriorate the correct estimations of statistical data from client scores. To mitigate the problem, some methods propose the suppresion of those scores which are far from the estimated scores mean. The tecnique results in a 'hard decision' that can produce errors especially when the number of scores is low. We propose here to take a 'softer decision' and weight scores according to their distance to the estimated scores mean. The Polycost and the BioTech databases have been used to show the effectiveness of the proposed method.

1 Introduction

The speaker verification is the process of deciding whether a speaker corresponds to a known voice. In speaker verification, the individual identifies her/himself by means of a code, login, card... Then, the system verifies her/his identity. It is a 1:1 process and it can be done in real-time. The result of the whole process is a binary decision. An utterance is compared to the speaker model and it is considered as belonging to the speaker if the Log-Likelihood Ratio (LLR) –the score obtained from the comparison- surpasses a predefined threshold and rejected if not.

In order to compare two systems, it is common to use the Equal Error Rate (EER), obtained when the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal. However, in real applications, a specific FAR or FRR is usually required. In this case, it is necessary to tune the speaker thresholds to achieve the desired rates.

In a typical Speaker Verification (SV) application, the user enrolls the system by pronouncing some utterances in order to estimate a speaker model. The enrollment procedure is one of the most critical stages of a SV process. At the same time, it

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 81 – 91, 2005.

becomes essential to carry out a successful training process to obtain a good performance. The importance and sensitiveness of the process force us to pay special attention on it. Consequently, it is necessary to protect the enrollment procedure by giving the user some security mechanisms, like extra passwords or by providing a limited physical access. A general block diagram of an SV process can be found in Figure 1:

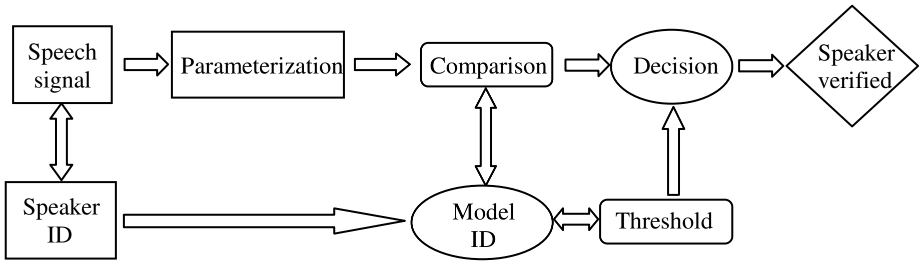


Fig. 1. Block diagram of a SV process

In real speaker verification applications, the speaker dependent thresholds should be estimated a priori, using the speech collected during the speaker models training. Besides, the client utterances must be used to train the model and also to estimate the threshold because data is scarce. It is not possible to use different utterances for both stages.

In development tasks, the threshold is usually set a posteriori. However, in real applications, the threshold must be set a priori. Furthermore, a speaker-dependent threshold can sometimes be used because it better reflects speaker peculiarities and intra-speaker variability than a speaker-independent threshold. The speaker dependent threshold estimation method uses to be a linear combination of mean, variance or standard deviation from clients and/or impostors.

Human-machine interaction can elicit some unexpected errors during training due to background noises, distortions or strange articulatory effects. Furthermore, the more training data available, the more robust model can be estimated. However, in real applications, one can normally afford very few enrolment sessions. In this context, the impact of those utterances affected by adverse conditions becomes more important in such cases where a great amount of data is not available. Score pruning (SP) [1,2,3] techniques suppress the effect of non-representative scores, removing them and contributing to a better estimation of means and variances in order to set the speaker dependent threshold. The main problem is that in a few cases the elimination of certain scores can produce unexpected errors in mean or variance estimation. In these cases, threshold estimation methods based on weighting the scores reduce the influence of the non-representative ones. The methods use a sigmoid function to weight the scores according to the distance from the scores to the estimated scores mean.

A theoretical approach of the state-of-the-art is reported on the next section. The weighting threshold estimation method is developed in section 3. The experimental setup, the description of the databases and the evaluation with empirical results are shown in section 4, followed by conclusions in section 5.

2 Theoretical Approach

Several approaches have been proposed to automatically estimate a priori speaker dependent thresholds. Conventional methods have faced the scarcity of data and the problem of an a priori decision, using client scores, impostor data, a speaker independent threshold or some combination of them. In [4], one can find an estimation of the threshold as a linear combination of impostor scores mean (μ_I) and standard deviation from impostors σ_I as follows:

$$\Theta = \alpha \ (\mu_I - \sigma_I) + \beta \tag{1}$$

where α and β should be empirically obtained.

Three more speaker dependent threshold estimation methods similar to (1) are introduced in (2), (3) and (4) [5, 6]:

$$\Theta = \mu_I + \alpha \ \sigma_I^2 \tag{2}$$

where $\hat{\sigma}_{\bar{X}}^2$ is the variance estimation of the impostor scores, and:

$$\Theta = \alpha \ \mu_I + (1 - \alpha) \ \mu_C \tag{3}$$

$$\Theta = \Theta_{SI} + \alpha \ (\mu_C - \mu_I) \tag{4}$$

where μ_c is the client scores mean, Θ_{SI} is the speaker independent threshold and α is a constant, different for every equation and empirically determined. Equation (4) is considered as a fine adjustment of a speaker independent threshold.

Another expression introduced in [1] encompasses some of these approaches:

$$\Theta = \alpha \ (\mu_I + \beta \ \sigma_I) + (1 - \alpha) \mu_C \tag{5}$$

where α and β are constants which have to be optimized from a pool of speakers.

An approach introduced by the authors in [2] uses only data from clients:

$$\Theta = \mu_C - \alpha \ \sigma_C \tag{6}$$

where μ_C is the client scores mean, σ_C is the standard deviation from clients and α is a constant empirically determined. Equation (6) is very similar to (2), but uses standard deviation instead of variance and the client mean instead of impostors mean.

Some other methods are based on FAR and FRR curves [7]. Speaker utterances used to train the model are also employed to obtain the FRR curve. On the other hand, a set of impostor utterances is used to obtain the FAR curve. The threshold is adjusted to equalize both curves.

There are also other approaches [8] based on the difficulty of obtaining impostor utterances which fit the client model, especially in phrase-prompted cases. In these cases, it is difficult to secure the whole phrase from impostors. The solution is to use the distribution of the ‘units’ of the phrase or utterance rather than the whole phrase. The units are obtained from other speakers or different databases.

On the other hand, it is worth noting that there are other methods which use different estimators for mean and variance. With the selection of a high percentage of frames and not all of them, those frames which are out of range of typical frame likelihood values are removed. In [9], two of these methods can be observed, classified according to the percentage of used frames. Instead of employing all frames, one of the estimators uses 95% most typical frames discarding 2.5% maximum and minimum frame likelihood values. An alternative is to use 95% best frames, removing 5% minimum values.

Normalization techniques [10] can also be used for threshold setting purposes. Some normalization techniques follow the Bayesian approach while other techniques standardise the impostor score distribution. Furthermore, some of them are speaker-centric and some others are impostor-centric.

Zero normalization (Znorm) [11, 12, 13] estimates mean and variance from a set of impostors as follows:

$$S_{M,norm} = \frac{S_M - \mu_I}{\sigma_I} \tag{7}$$

where S_M are the client scores, μ_I is the estimated mean from impostors and σ_I the estimated variance from impostors [14].

We should also mention another threshold normalization technique such as Test normalization (Tnorm) [13, 15], which uses impostor models instead of impostor speech utterances to estimate impostor score distribution. The incoming speech utterance is compared to the speaker model and to the impostor models. That is the difference with regard to Znorm. Tnorm also follows the equation (7).

Tnorm has to be performed on-line, during testing. It can be considered as a test-dependent normalization technique while Znorm is considered as a speaker-dependent one. In both cases, the use of variance provides a good approximation for the impostor distribution.

Furthermore, Tnorm has the advantage of matching between test and normalization because the same utterances are used for both purposes. That is not the case for Znorm.

Finally, we can also consider Handset normalization (Hnorm) [16, 17, 18]. It is a variant of Znorm that normalizes scores according to the handset. This normalization is very important especially in those cases where there is a mismatch between training and testing.

Since handset information is not provided for each speaker utterance, a maximum likelihood classifier is implemented with a GMM for each handset [17]. With this classifier, we decide which handset is related to the speaker utterance and we obtain mean and variance parameters from impostor utterances. The normalization can be applied as follows:

$$S_{M,norm} = \frac{S_M - \mu_I(handset)}{\sigma_I(handset)} \tag{8}$$

where μ_I and σ_I are respectively the mean and variance obtained from the speaker model against impostor utterances recorded with the same handset type, and S_M are the client scores.

3 Threshold Estimation Based on Weighting Scores

A threshold estimation method that weights the scores according to the distance d_n from the score to the mean is introduced [19] in this section. It is considered that a score which is far from the estimated mean comes from a non-representative utterance of the speaker. The weighting factor w_n is a parameter of a sigmoid function and it is used here because it distributes the scores in a nonlinear way according to their proximity to the estimated mean. The expression of w_n is:

$$w_n = \frac{1}{1 + e^{-C d_n}} \tag{9}$$

where w_n is the weight for the utterance n, d_n is the distance from the score to the mean and C is a constant empirically determined in our case.

The distance d_n is defined as:

$$d_n = \left| s_n - \mu_s \right| \tag{10}$$

where s_n are the scores and μ_s is the estimated scores mean.

The constant C defines the shape of the sigmoid function and it is used to tune the weight for the sigmoid function defined in Equation (9). A positive C will provide increasing weights with the distance while a negative C will give decreasing values. A typical sigmoid function, with C=1 is shown in Figure 2:

The average score is obtained as follows:

$$s_T = \frac{\sum_{n=1}^{N} w_n s_n}{\sum_{n=1}^{N} w_n} \tag{11}$$

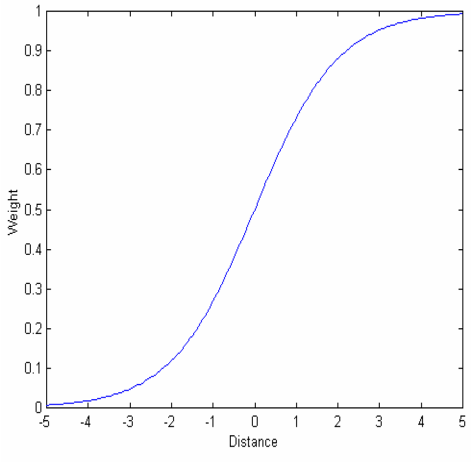


Fig. 2. Sigmoid function

where w_n is the weight for the utterance n defined in (9), s_n are the scores and s_T is the final score.

The standard deviation is also weighted in the same way as the mean. This method is called Total Score Weighting (T-SW).

On the other hand, it is possible to assign weights different from 1.0 only to a certain percentage of scores –the least representative- and not to all of them. This method is called Partial Score Weighting (P-SW). Normally, the farthest scores have in this case a weight different from 1.0.

4 Experiments

4.1 The Polycost Database

The Polycost database has been used for the experiments in this work. It was recorded by the participants of the COST250 Project. It is a telephone speech database with 134 speakers, 74 male and 60 female. The 85% of the speakers are between 20 and 35 years old. Almost each speaker has between 6 and 15 sessions of one minute of speech. Most speakers were recorded during 2-3 months, in English and in their mother tongue. Calls were made from the Public Switched Telephone Network (PSTN).

Each session contains 14 items: 4 repetitions of a 7-digit client code, five 10-digit sequences, 2 fixed sentences, 1 international phone number and 2 more items of spontaneous speech in speaker's mother tongue. For our experiments, we will use only digit utterances in English.

4.2 The BioTech Database

One of the databases used in this work has been recorded –among others- by the author and has been especially designed for speaker recognition. It is called the BioTech database and it belongs to the company Biometric Technologies, S.L. It includes landline and mobile telephone sessions. A total number of 184 speakers were recorded by phone, 106 male and 78 female. It is a multi-session database in Spanish, with 520 calls from the Public Switched Telephone Network (PSTN) and 328 from mobile telephones. One hundred speakers have at least 5 or more sessions. The average number of sessions per speaker is 4.55. The average time between sessions per speaker is 11.48 days.

Each session includes:

- different sequences of 8-digit numbers, repeated twice. They were the Spanish personal identification number and that number the other way round. There were also two more digits: 45327086 and 37159268.
- different sequences of 4-digit numbers, repeated twice. They were one random number and the fixed number 9014.
- different isolated words.
- different sentences.
- 1 minute long read paragraph.

- 1 minute of spontaneous speech, suggesting to talk about something related to what the user could see around, what (s)he had done at the weekend, the latest book read or the latest film seen.

4.3 Setup

In our experiments, utterances are processed in 25 ms frames, Hamming windowed and pre-emphasized. The feature set is formed by 12th order Mel-Frequency Cepstral Coefficients (MFCC) and the normalized log energy. Delta and delta-delta parameters are computed to form a 39-dimensional vector for each frame. Cepstral Mean Subtraction (CMS) is also applied.

Left-to-right HMM models with 2 states per phoneme and 1 mixture component per state are obtained for each digit. Client and world models have the same topology. The silence model is a GMM with 128 Gaussians. Both world and silence models are estimated from a subset of their respective databases.

The speaker verification is performed in combination with a speech recognizer for connected digits recognition. During enrolment, those utterances catalogued as "no voice" are discarded. This ensures a minimum quality for the threshold setting.

In the experiments with the BioTech database, clients have a minimum of 5 sessions. It yields 100 clients. We used 4 sessions for enrolment –or three sessions in some cases- and the rest of sessions to perform client tests. Speakers with more than one session and less than 5 sessions are used as impostors. 4- and 8-digit utterances are employed for enrolment and 8-digit for testing. Verbal information verification [20] is applied as a filter to remove low quality utterances. The total number of training utterances per speaker goes from 8 to 48. The exact number depends on the number of utterances discarded by the speech recognizer. During test, the speech recognizer discards those digits with a low probability and selects utterances which have exactly 8 digits. A total number of 20633 tests have been performed for the BioTech database, 1719 client tests and 18914 impostor tests.

It is worth noting that land-line and mobile telephone sessions are used indistinctly to train or test. This factor increases the error rates.

On the other hand, only digit utterances are used to perform tests with the Polycost database. After using a digit speech recognizer, those speakers with at least 40 utterances where considered as clients. That yields 99 clients, 56 male and 43 female. Furthermore, the speakers with a number of recognized utterances between 25 and 40 are treated as impostors. If the number of utterances does not reach 25, those speakers are used to train the world model. We use 40 utterances to train every client model.

In the experiments with the Polycost database, 43417 tests were performed, 2926 client tests and 40491 impostor tests. All the utterances come from landline phones in contrast with the utterances that belong to the BioTech database.

4.4 Results

In this section, the experiments show the performance of the threshold estimation methods proposed here. The following table shows a comparison of the EER for threshold estimation methods with client data only, without impostors and for the baseline Speaker-Dependent Threshold (SDT) method defined in Equation (6).

As it can be seen in Table 1, the T-SW method performs better than the baseline and even than the SP method. The P-SW performs better than the baseline too, but not than the SP. The results shown here correspond to the weighting of the scores which

Table 1. Comparison of threshold estimation methods in terms of Equal Error Rate

SDT	Baseline	SP	T-SW	P-SW
EER (%)	5.89	3.21	3.03	3.73

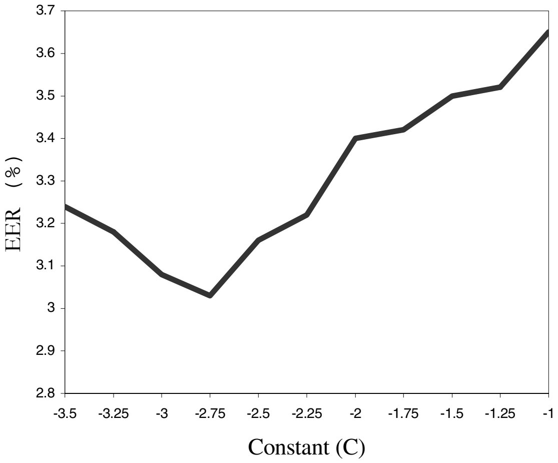


Fig. 2. Evolution of the EER with the variation of C

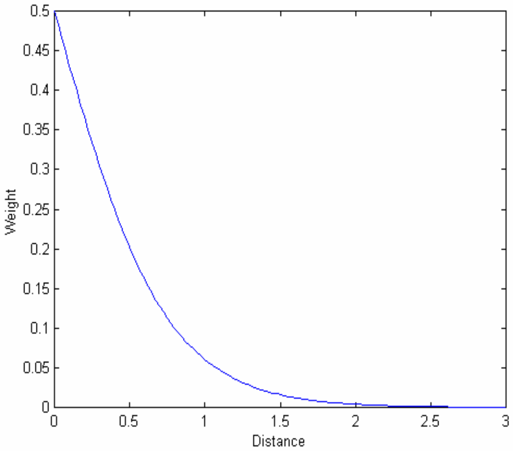


Fig. 3. Variation of the weight (w_n) with respecto to the distance (d_n) between the scores and the scores mean

distance to the mean is bigger than the 10% of the most distant score. It has been found that the minimum EER is secured when every one of the scores is weighted. It means that the optimal case for the P-SW method is the T-SW method.

In Figure 2, we can see the EER with respect to the constant C. It has been shown that the system performs better for a C = -2.75.

Figure 3 shows the function of the distance and the weight for the best C = -2.75. The weight exponentially decreases with the distance.

Table 2 shows the experiments with speaker-dependent thresholds using only data from clients following Equation (6).

The best EER is obtained for the Score Pruning (SP) method. The T-SW performs slightly worse and P-SW is the worst method. SP and SW methods improve the error rates with regard to the baseline. Results are given for a constant C = -3.0.

In Figure 4, the best EER is obtained for C = -3. This value is very similar to the one obtained for the BioTech database (C = -2.75).

Table 2. Comparison of threshold estimation methods for the Polycost database

SDT	Baseline	SP	T-SW	P-SW
EER (%)	1.70	0.91	0.93	1.08

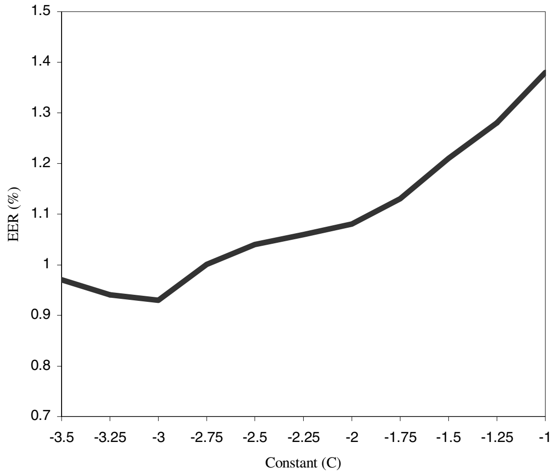
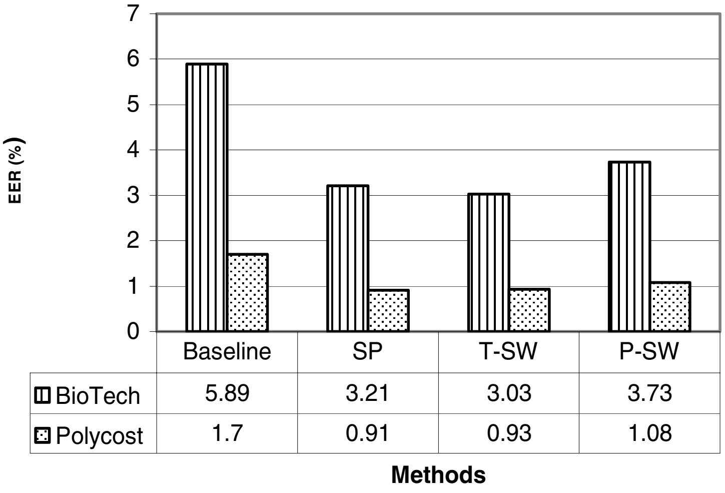


Fig. 4. Evolution of the EER with the variation of C

The comparison of the results obtained with both databases can be seen in Figure 5. First of all, EERs are lower for the Polycost database, mainly due to the fact that utterances are recorded from the PSTN while in the BioTech database calls come from the landline phones and the mobile phones. Furthermore, in the experiments with the BioTech database, some clients are trained for example with utterances recorded from fixed-line phones and then tested with utterances from mobile phones and this random use of sessions decreases performance.



	Baseline	SP	T-SW	P-SW
▥ BioTech	5.89	3.21	3.03	3.73
▨ Polycost	1.7	0.91	0.93	1.08

Fig. 5. Comparison of EERs obtained for the BioTech and the Polycost databases

On the other hand, the improvement obtained with SP and SW methods is larger in experiments with the Polycost database where it almost reaches the 50%.

Otherwise, SP method gives an EER similar to the T-SW method in experiments with the Polycost database. On the contrary, T-SW method performs clearly better than SP method in the experiments with the BioTech database. The P-SW method is the method with the worst performance in both cases.

5 Conclusions

The automatic estimation of speaker dependent thresholds has revealed as a key factor in speaker verification enrolment. Threshold estimation methods mainly deal with the sparseness of data and the difficulty of obtaining data from impostors in real-time applications. These methods are currently a linear combination of the estimation of means and variances from clients and/or impostor scores. When we have only a few utterances to create the model, the right estimation of means and variances from client scores becomes a real challenge.

Although the SP methods try to mitigate main problems by removing the outliers, another problem arises when only a few scores are available. In these cases, the suppression of some scores worsens estimations. For this reason, weighting threshold methods proposed here use the whole set of scores but weighting them in a nonlinear way according to the distance to the estimated mean. Weighting threshold estimation methods based on a nonlinear function improve the baseline speaker dependent threshold estimation methods when using data from clients only. The T-SW method is even more effective than the SP ones in the experiments with the BioTech database, where there is often a mismatched handset between training and testing. On the contrary, with the Polycost database, where the same handset (landline network) is used, both of them perform very similar.

References

1. Chen, K.: Towards Better Making a Decision in Speaker Verification. Pattern Recognition, Vol. 36 (2003) 329-346
2. Saeta, J.R., Hernando, J.: Automatic Estimation of A Priori Speaker Dependent Thresholds in Speaker Verification. In: Proceedings 4th International Conference in Audio- and Video-based Biometric Person Authentication (AVBPA). Lecture Notes in Computer Science. Springer-Verlag (2003) 70-77
3. Saeta, J.R., Hernando, J.: On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation. 2004: A Speaker Odyssey, The Speaker Recognition Workshop (2004) 215-218
4. Furui, S.: Cepstral Analysis for Automatic Speaker Verification. IEEE Trans. Speech and Audio Proc., vol. 29(2) (1981) 254-272
5. Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Pierrot, J.B., Blomberg, M., Bimbot, F.: Techniques for A Priori Decision Threshold Estimation in Speaker Verification. In: Proceedings RLA2C (1998) 89-92
6. Pierrot, J.B., Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Blomberg, M., Bimbot, F.: A Comparison of A Priori Threshold Setting Procedures for Speaker Verification in the CAVE Project. In: Proceedings ICASSP (1998) 125-128
7. Zhang, W.D., Yiu, K.K., Mak, M.W., Li, C.K., He, M.X.: A Priori Threshold Determination for Phrase-Prompted Speaker Verification. In: Proceedings Eurospeech (1999) 1203-1206
8. Surendran, A.C., Lee, C.H.: A Priori Threshold Selection for Fixed Vocabulary Speaker Verification Systems. In: Proceedings ICSLP vol. II (2000) 246-249
9. Bimbot, F., Genoud, D.: Likelihood Ratio Adjustment for the Compensation of Model Mismatch in Speaker Verification. In: Proceedings 2001: A Speaker Odyssey, The Speaker Recognition Workshop (2001) 73-76
10. Gravier, G. and Chollet, G.: Comparison of Normalization Techniques for Speaker Verification. In: Proceedings RLA2C (1998) 97-100
11. Auckentaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Verification Systems. Digital Signal Processing, Vol. 10 (2000) 42-54
12. Bimbot, F., Bonastre, F.J., Fredouille, C., Gravier, G., Magrin, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska, D., Reynolds, D.: A Tutorial on Text-Independent Speaker Verification. In: Proceedings Eusipco (2004) 430-451
13. Mirghafori, N., Heck, L.: An Adaptive Speaker Verification System with Speaker Dependent A Priori Decision Thresholds. In: Proceedings ICSLP (2002) 589-592
14. Navratil, J., Ramaswamy, G.N.: The Awe and Mystery of T-norm. In: Proceedings Eurospeech (2003) 2009-2012
15. Reynolds, D.: The Effect of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus. In: Proceedings ICASSP (1996) 113-116, 1996
16. Reynolds, D.A.: Comparison of Background Normalization Methods for Text-Independent Speaker Verification. In: Proceedings Eurospeech (1997) 963-966
17. Heck, L.P., Weintraub, M.: Handset Dependent Background Models for Robust Text-Independent Speaker Recognition. In: Proceedings ICASSP (1997) 1071-1074
18. Saeta, J.R., Hernando, J.: New Speaker-Dependent Threshold Estimation Method in Speaker Verification based on Weighting Scores. In Proceedings of the 3th Internacional Conference on Non-Linear Speech Processing (NoLisp) (2005) 34-41
19. Li, Q., Juang, B.H., Zhou, Q., Lee, C.H.: Verbal Information Verification. In: Proceedings Eurospeech (1997) 839-842