INCREASING ROBUSTNESS TO TRAINING-TEST MISMATCH IN SPEAKER VERIFICATION THROUGH SHALLOW SOURCE MODELLING

M. Zamalloa†‡, L. J. Rodriguez-Fuentes†, M. Penagarikano†, G. Bordel†, J. P. Uribe‡

†Grupo de Trabajo en Tecnologías del Software, DEE, ZTF/FCT, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, SPAIN

‡Ikerlan – Technological Research Centre
Paseo J.M. Arizmendiarrieta 2, 20500 Arrasate-Mondragón, SPAIN

phone: +34 946012716, fax: +34 946013500, email: luisjavier.rodriguez@ehu.es
web: http://gtts.ehu.es/TWiki/bin/view (in Spanish)

ABSTRACT

Speaker verification is usually performed by comparing the likelihood score of the target speaker model to the likelihood score of an universal background model (UBM), and then applying a suitable threshold. For the UBM to be effective, it must be estimated from a large number of speakers. However, it is not always possible to gather enough data to estimate a robust UBM, and the verification performance may degrade if impostors, or whatever sources that generate the input signals, were not suitably modelled by the UBM. In this work, a new normalization technique is proposed, based on a shallow source model (SSM) estimated from the input utterance. A linear combination of the likelihood scores of the SSM and the UBM is used to normalize the speaker score. Speaker verification experiments were carried out on a clean-speech dataset including 204 speakers. Also, a sizeable amount of noisy, speech and non-speech signals was used to test the robustness to large training-test mismatch. Three normalization techniques were tested: UBM, smoothed UBM and the proposed combination of UBM and SSM. This latter approach yielded the best performance. The difference in performance was specially significant in the large training-test mismatch condition.

1. INTRODUCTION

Speaker verification consists of deciding whether an input utterance X was *actually* produced by the claimed speaker or not. This task naturally arises in environments requiring biometric authentication of any potential user. In fact, using the voice as a biometric is probably the most natural way of authenticating people. Another interesting application is speaker tracking in broadcast news: the audio signal is segmented into homogeneous sections (usually speaker turns), which are then automatically labelled either with the name of a target speaker or with the name of a default category corresponding to unknown speakers and other sources (music, noise, etc.).

Though speaker characteristics are reflected at many levels (acoustic, phonetic, phonological, prosodic, syntactic or even pragmatic), most systems take into account only the physiological information conveyed by the acoustic parameters extracted from the speech signal, and use an acoustic model to gather the statistics of the power spectrum specific to each speaker. It is assumed that speech data are available for the target speaker, so that an acoustic model λ can be estimated. An input utterance is represented by a sequence of acoustic vectors $X = \{x_1, x_2, \ldots, x_T\}$. From a practical point of view, the verification task can be reduced to computing the posterior probability $P(\lambda|X)$ and comparing it to a fixed threshold τ. The input utterance X is accepted as coming from the claimed speaker if the *verification inequality* $P(\lambda|X) > \tau$ holds; otherwise, it is rejected. The decision threshold τ can be heuristically adjusted to get a suitable trade-off between false acceptance and false rejection errors.

Applying the Bayes rule, the verification inequality can be written as follows:

$$\frac{P(X|\lambda)P(\lambda)}{P(X)} > \tau$$

Since the prior probability $P(\lambda)$ does not depend on X, and taking logarithms, the verification inequality can be rewritten in terms of a *log-likelihood ratio*:

$$LLR_1(X) = \log P(X|\lambda) - \log P(X) > \tau_1 \qquad (1)$$

In most cases, the conditional probability $P(X|\lambda)$ is represented by means of a *Gaussian Mixture Model* (GMM) [7]. The normalizing term $P(X)$ is the probability of the input utterance X. To compute $P(X)$, a *source model* λ_S is needed which accounts for all the potential input utterances, so that $P(X) = P(X|\lambda_S)$. If the source was known, a large, diverse and balanced database of audio samples coming from that source might be used to estimate a suitable acoustic model λ_S.

From the point of view of decision theory, the verification task consists of deciding between the H_0 hypothesis (X belongs to the claimed speaker) or the alternative H_1 hypothesis (X does not belong to the claimed speaker). This may be accomplished by computing the posterior probabilities $P(H_0|X)$ and $P(H_1|X)$ and accepting H_0 if the likelihood ratio is greater than a given threshold:

$$\frac{P(H_0|X)}{P(H_1|X)} > \tau'$$

Since the prior probabilities $P(H_0)$ and $P(H_1)$ do not depend on X, applying the Bayes rule and taking logarithms, a log-likelihood ratio is obtained:

$$LLR_2(X) = \log P(X|H_0) - \log P(X|H_1) > \tau_2 \qquad (2)$$

In this case, the speaker score $P(X|H_0)$ is normalized by the score of an *impostor model*, since it represents the alternative to the claimed speaker (i.e. impostors). Note that Eqs.

1 and 2 differ only in the normalizing term. In fact, $P(X|\lambda) = P(X|H_0)$ and $P(X) = P(X|H_0)P(H_0) + P(X|H_1)P(H_1)$.

Here, we follow the approach given by Eq. 1. The source model $P(X)$ is usually called *background model*, since it provides acoustic coverage for a wide range of input utterances. Various alternatives have been proposed in the literature to define a suitable source model. One of them consists of defining the source as a combination of known sources: a *cohort of background speakers* selected according to a given criterion of closeness, remoteness, competitiveness or the like, with regard to the target speaker [9]. An acoustic model is estimated for each source, and the likelihood score $P(X)$ is computed as a function (usually the arithmetic mean) of the likelihood scores of the potential sources. Two issues arise with this approach: (1) a suitable cohort of background speakers must be selected and combined for each target speaker; and (2) it is not easy to cover all the potential input utterances with just a few background speakers.

The most common approach to modelling the source consists of using a large pool of speakers to train a single speaker-independent model, called *Universal Background Model* (UBM), usually a GMM with a large number of components [6], designed to match the statistics of any potential input utterance. The UBM approach has several advantages: (1) a single model is used to normalize the likelihood scores of all the target speakers; (2) it provides universal acoustic coverage for speech signals; and (3) it can be used as the basis for estimating speaker models through Bayesian adaptation, thus yielding more robust speaker models. However, it is not always possible to gather enough data to estimate such a robust UBM. Note also that, depending on the application (for example, speaker tracking), the source could be non-human (music, noise, etc.). In this case, and whenever the source that generates the input utterance was not suitably modelled by the UBM, neither the speaker model nor the UBM would cover the input utterance, the log-likelihood ratio would not be reliable and the verification performance would degrade.

In this work, we aim to improve the source model given by the UBM. Instead of taking as reference only an estimation of what input signals should be like (the UBM), we also take as reference an estimation of the source based on the input signal. We estimate an acoustic model of the source that generates the input utterance, that we call *Shallow Source Model* (SSM), and then use a linear combination of the likelihood scores of the UBM and the SSM to normalize the speaker score. This approach solves the issue of coverage, since the SSM just attempts to model the source that generates the input utterance. Raw SSM normalization (without UBM) was originally applied to speaker tracking in broadcast news [8]. More recently, the mixed UBM-SSM approach has been successfully applied to open-set speaker identification [13]. In this paper we go more deeply into that line of research, by changing the focus to speaker verification, using more up-to-date speaker models (MAP-adapted from the UBM) and including performance comparison to a similar approach by other authors.

Few alternatives to the UBM, such as the one presented in this paper, can be found in the literature. The same principle of using a *weak* or *low acoustic resolution* model to normalize speaker scores, instead of a large high-resolution UBM, has been previously applied in text-dependent speaker verification by Siohan et al. [10], and in text-independent speaker verification by Tran [11]. In both cases, authors try to circumvent the need for a large speaker-independent database by exploiting the enrollment data of target speakers in a smart way, but input utterances are not used in any way. On the other hand, Hsu, Yu and Yang [5] estimate an acoustic model from the input utterance and take it as reference to make a decision. However, the verification procedure they propose, based on the tolerance interval analysis, use speaker samples instead of speaker models. Finally, Tran and Wagner [12] present experimental results supporting the claim that a sizeable number of false acceptances can be avoided by smoothing the UBM likelihood score with a constant membership value ε. This constant plays the same role as the SSM, but in a blind way, since no information is extracted from the input utterance.

The rest of the paper is organized as follows. Section 2 briefly describes the SSM and presents a way of combining the UBM and the SSM to get a more robust source model. Section 3 gives details about the datasets and the baseline system used in the speaker verification experiments. Results are presented and discussed in Section 4. Finally, conclusions are given in Section 5.

2. THE SHALLOW SOURCE MODEL

This approach consists of estimating a source model λ_X from the input utterance X, computing the score $P(X|\lambda_X)$ and using it to normalize the speaker score. Since X is usually short (2-10 seconds), a low-order (*shallow*) model must be defined, to allow robust estimates and avoid overtraining. Note that we do not aim to model the input utterance but the source (for instance, the speaker, but also other kind of sources), and using too many mixtures would model utterance-specific variations instead of source-generic features. In summary, a very simple and shallow model (currently, a GMM), which we call *Shallow Source Model* (SSM), is estimated to model the source.

If the SSM λ_X (estimated from the input utterance X) was a *perfect* source model, then, for any speaker model λ (estimated from independent training samples), it should be:

$$P(X|\lambda_X) > P(X|\lambda) \tag{3}$$

In these conditions, the *log-likelihood ratio*:

$$LLR(X) = \log P(X|\lambda) - \log P(X|\lambda_X)$$

would be always negative or zero, and it would be zero only in the case the speaker model λ perfectly matched the source model λ_X. Clearly, in this latter case the input utterance should be accepted, but the same decision should be made if the log-likelihood ratio was close enough to zero. In fact, using the source model score to normalize the speaker model score gives a measure of *how well the speaker model approximates the source model*: if the log-likelihood ratio was greater than a given threshold, then the input utterance would be accepted as belonging to the claimed speaker; otherwise, it would be rejected.

In practice, however, λ_X is not a perfect but a shallow source model and the inequality 3 does not hold. Speaker models are acoustically rich GMMs, trained on much more data than the SSM, so some of them may cover the input utterance better than the SSM. Nevertheless, the likelihood score of the SSM may still be taken as a reference to normalize speaker scores, and a suitable threshold applied to make

a decision. The same interpretation given above holds in this case: the SSM provides a reference to measure how well the speaker model approximates the source. Moreover, if the input utterance X was not suitably covered by speaker models (because it comes from an impostor, or from a non-human source), the SSM would still *guarantee* a minimum acoustic coverage (playing the same role as the constant ε proposed by Tran and Wagner in [12]). Its likelihood score would be higher than that of the speaker model, and X would be reliably classified as an impostor utterance.

2.1 Combining the UBM and the SSM

The SSM approach solves the issue of acoustic coverage and does not need lots of data as the UBM does, but SSM parameter estimates may be highly influenced by utterance-specific variations so that the SSM would not be robustly modelling the source. In fact, previous experimentation has shown that the UBM clearly outperforms the SSM in this kind of tasks [13].

To overcome the coverage issue of the UBM and the robustness issue of the SSM, the background model may be estimated by Bayesian adaptation of the UBM to the input utterance (see [2]). However, this approach takes more computation than simply estimating the SSM. Alternatively, the likelihood score of the input utterance can be approximated by a suitable linear combination of the likelihood scores of the UBM and the SSM, as follows:

$$P(X) \approx \alpha P(X|\lambda_{\text{UBM}}) + (1-\alpha)P(X|\lambda_X) \qquad (4)$$

where α is a heuristically fixed mixing factor (i.e. the optimal value of α is that yielding the best performance on the test set). Equation 4 expresses the assumption that X is generated either by the UBM, which robustly accounts for a wide range of speakers, or the SSM, which provides a weak estimation of whatever other sources. From this point of view, the SSM guarantees full acoustic coverage of the input utterances. Finally, this approach can be applied to any database, since it only requires estimating an SSM for each input utterance, and then combining the SSM score with that of the UBM. There is only an issue, related to the mixing factor α, whose optimal value should be fixed by optimizing the performance on a validation dataset.

3. EXPERIMENTAL SETUP

3.1 Datasets

A phonetically balanced database in Castilian Spanish, called *Albayzín* [3], recorded at 16 kHz in laboratory conditions, was used in the experiments. The database contains 204 speakers, each contributing at least 25 read utterances, and each utterance lasting an average of 3.55 seconds. Albayzín was originally designed to train acoustic models for speech recognition and may be considered similar in characteristics to TIMIT. Nowadays, it is the most widely used acoustic-phonetic database in Spanish.

For the experiments presented in this paper, a gender-balanced set of 34 target speakers and a gender-balanced set of 68 impostors were considered, the remaining 102 being used as background speakers. Three disjoint sets of utterances were considered: (1) the *training set*, consisting of 10 utterances from each target speaker, was used to estimate speaker models; (2) the *background set*, consisting of 25 utterances from each background speaker, was used to estimate the UBM; and (3) the *test set*, consisting of 15 utterances from each target speaker and 15 utterances from each impostor, was used to evaluate the performance of the speaker verification systems. So, the dataset consists of 340 training utterances, 2550 background utterances and 1530 test utterances (of which 510 correspond to target speakers and 1020 to impostors).

Besides Albayzín, a separate dataset, called *Mismatched*, was used to test the robustness of speaker verification systems to signals not suitably covered by the training material. *Mismatched* was designed to match the size and structure of the test set of Albayzín. It consists of 1920 utterances (each lasting 3 seconds) resampled at 16 kHz. The corpus is divided into three different subcorpora: (1) *Music*, consisting of 576 song fragments taken at random from a song database; (2) *Telephone*, consisting of 640 spontaneous speech fragments taken at random from *Dihana* [1], a database of human-computer dialogues recorded at 8 kHz through telephone lines; and (3) *WWW*, consisting of 664 audio fragments (most of them including speech) taken at random from the internet.

3.2 The baseline system

A state-of-the-art GMM/UBM speaker verification system was applied in the experiments. Speaker models were estimated by MAP adaptation of the UBM to the training dataset of each speaker [6]. The UBM was defined as a 1024-component GMM, whose parameters were estimated by Maximum Likelihood from the set of background speakers, using the EM algorithm and starting from random values. Mel Frequency Cepstral Coefficients (MFCC) with mean normalization were used as acoustic features. The frame energy was also computed, yielding a 13-dimensional feature vector.

4. RESULTS AND DISCUSSION

4.1 Speaker verification on a test set fully covered by training data

First, speaker verification experiments were run on the test set of Albayzín, using the training and background datasets to estimate the speaker models and the UBM, respectively. Besides the UBM, two additional normalization techniques were tested: the smoothed UBM proposed by Tran and Wagner [12] and the linear combination of the UBM and SSM likelihoods proposed in this paper.

A preliminary series of experiments was run to determine the optimal size of the GMM used to represent the source in the SSM approach. It was expected to be a low value, since a large GMM would be less robust and would model utterance-specific features. The best performance was obtained for a SSM with 4 mixture components.

To compare the performance of speaker verification systems, results are presented in the form of DET (*Detection Error Trade-off*) curves. DET curves are generated by using the DET-Curve Plotting software provided by NIST [4]. As shown in Figure 1, a suitable linear combination of the UBM and SSM likelihoods (mixing factor $\alpha = 0.97$) slightly improves the performance of the UBM (the EER decreasing from 0.9% to 0.8%), whereas the smoothed UBM proposed by Tran and Wagner (background membership constant $\varepsilon = 0.005$) does not yield any improvement.

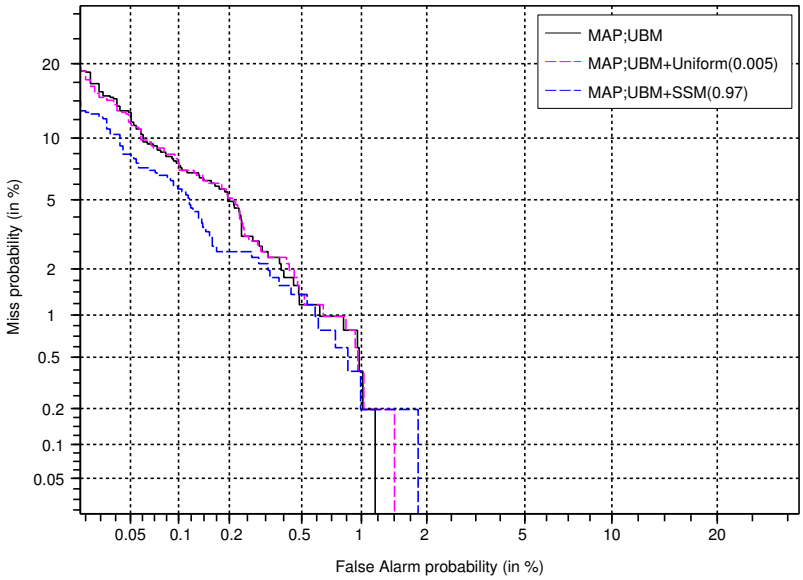

Figure 1: DET curves for three speaker verification systems using MAP adapted speaker models and three normalization methods: UBM, smoothed UBM ($\varepsilon = 0.005$) and a suitable combination of the UBM and SSM likelihoods ($\alpha = 0.97$). Experiments were carried out on the test set of Albayzín, which was suitably covered by the UBM.

4.2 Speaker verification on a test set partially covered by training data

A second series of experiments was run to evaluate the robustness of speaker verification systems to signals not covered by training data. The speaker and background models were the same applied in the experiments described in Section 4.1. The test set was extended with the *Mismatched* dataset. It comprised 3450 utterances, 510 coming from target speakers, 1020 from impostor speakers in matched conditions and 1920 from impostor speakers and other sources in uncovered/mismatched conditions (see Section 3.1 for details). In this way we tried to simulate the situation where a relatively large amount of signals in mismatched conditions must be processed. That is the case of speaker tracking applications, where input signals are segmented into acoustically homogeneous regions, and these assigned either to a target speaker or to a generic unknown source.

For each target speaker, the test set comprises 15 utterances coming from the target speaker and 3435 coming from impostor speakers and other sources not suitably covered by training data. This amounts to $34 \cdot 3450 = 117300$ verification tests, but only 510 of them are used to compute miss rates, whereas 116790 are used to compute false alarm rates, which makes miss rates less reliable than false alarm rates. Figure 2 shows the DET curves resulting from speaker verification experiments on the extended test set using the three normalization approaches described above.

Again, the linear combination of the UBM and SSM likelihoods ($\alpha = 0.97$) consistently outperformed UBM, whereas the smoothing procedure proposed by Tran and Wagner ($\varepsilon = 0.005$) yielded no significant improvement. Differences in performance were larger at the high-thresholds end of the DET curve (10% vs. 19% miss rate at 0% false alarm rate), and became almost null at the low-thresholds end (ranging between 1.6% and 1.7% false alarm rate at 0% miss rate).

The linear combination of the UBM and SSM likelihoods provided more significant improvements with regard to the UBM when applied to the extended test set (0.4% vs. 0.6% EER) than when applied to the baseline test set (0.8% vs. 0.9% EER). This means that the main contribution of the SSM is helping reject utterances in uncovered/mismatched conditions. As a result, the false alarm rates in Figure 2 are lower than those in Figure 1 (e.g. moving from 0.9% to 0.4% at 0.5% miss rate). The UBM also succeeds in rejecting a sizeable amount of utterances in mismatched conditions, but not so much as the combination of UBM and SSM. As a result, the false alarm rates of UBM in Figure 2 do also decrease with regard to those in Figure 1, but to a lesser extent (e.g. moving from around 1% to 0.65% EER at 0.5% miss rate).

5. CONCLUSIONS

In this paper a new approach to the issue of normalizing speaker scores in speaker verification is presented which aims to improve the robustness to training-test mismatch. Besides modelling a wide range of potential speakers by estimating a Universal Background Model (UBM), a low-order GMM —which we call *Shallow Source Model* (SSM)— is estimated from the input utterance. Then, a suitable linear combination of the UBM and SSM likelihoods is used to normalize the speaker score. This approach solves the issue of acoustic coverage, because it includes a model of the source that generates the input utterance. On the other hand, estimating the SSM and computing its likelihood do not increase significantly the computational cost of speaker verification.

The proposed approach has been compared to the UBM and a smoothed version of the UBM —which is reported to be more robust to signals not covered by training data— in two series of speaker verification experiments: (1) on a test set fully covered by training data, and (2) on a test set including a sizeable amount of signals not covered by training data. In both cases, the proposed approach yielded better results than the UBM, whereas the smoothed UBM did not yield significant improvements with regard to the UBM. The im-

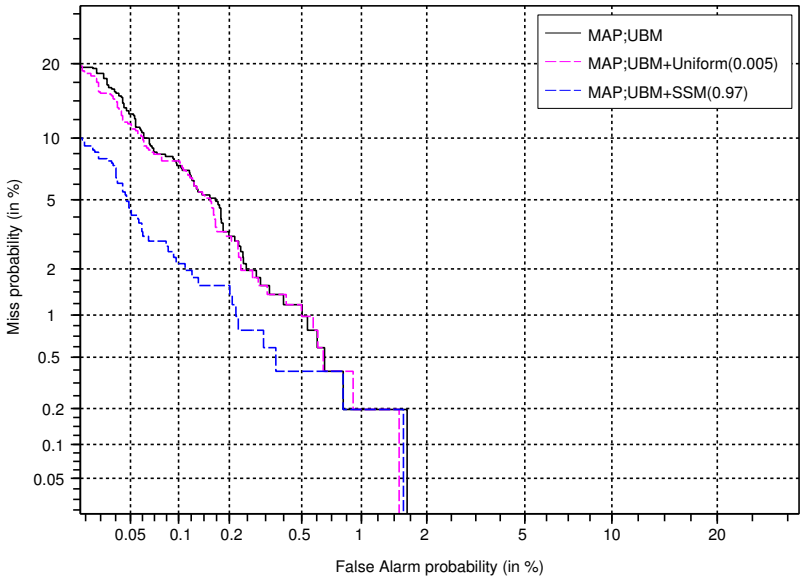

Figure 2: DET curves for three speaker verification systems using MAP adapted speaker models and three normalization methods: UBM, smoothed UBM ($\varepsilon = 0.005$) and a suitable combination of the UBM and SSM likelihoods ($\alpha = 0.97$). In this case, the test set of Albayzín was augmented with noisy, speech and non-speech signals not suitably covered by the UBM.

provement provided by the SSM was more noticeable when dealing with signals not matching training data. Though the UBM successfully rejected a sizeable amount of utterances not covered by training data, the use of SSM helped reject even more of them, leading to lower false alarm rates.

Current work includes testing the mixed UBM-SSM approach in more realistic conditions, by using speaker databases recorded on different channels and different sessions. Also, the mixing factor α will be fixed by optimizing the performance on a validation dataset.

6. ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish MEC, under Plan Nacional de I+D+i, project TSI2006-14250-C02-01; the Government of the Basque Country, under program SAIOTEK, projects S-PE06UN48, S-PE06IK01, S-PE07UN43 and S-PE07IK03; and the University of the Basque Country, under project EHU06/96.

REFERENCES

[1] N. Alcocer, M. J. Castro, I. Galiano, R. Granel, S. Grau, and G. D. Adquisición de un Corpus de Diálogo: DIHANA. In *III Jornadas en Tecnología del Habla (in Spanish)*, pages 131–134, 2004.

[2] H. Aronowitz, D. Burshtein, and A. Amihood. A Session-GMM Generative Model Using Test Utterance Gaussian Mixture Modeling for Speaker Verification. In *Proceedings of ICASSP*, volume 1, pages 733–736, 2005.

[3] F. Casacuberta, R. García, J. Llisterri, C. Nadeu, J. M. Pardo, and A. Rubio. Development of Spanish Corpora for Speech Research (Albayzín). In *Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods*, pages 26–28, 1991.

[4] DET-Curve Plotting software for use with MATLAB. www.nist.gov/speech/tools/DETware_v2.1.targz.htm.

[5] C.-N. Hsu, H.-C. Yu, and B.-H. Yang. Speaker Verification Without Background Speaker Models. In *Proceedings of ICASSP*, volume II, pages 233–236, 2003.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, January/April/July 2000.

[7] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.

[8] L. J. Rodríguez, M. Peñagarikano, and G. Bordel. A Simple But Effective Approach to Speaker Tracking in Broadcast News. In *J. Martí et al. (Eds.): IbPRIA 2007, Proceedings, LNCS 4478*, pages 48–55, 2007.

[9] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong. The Use of Cohort Normalized Scores for Speaker Verification. In *Proceedings of ICSLP*, pages 599–602, 1992.

[10] O. Siohan, C. H. Lee, A. C. Surendran, and Q. Li. Background Model Design for Flexible and Portable Speaker Verification Systems. In *Proceedings of ICASSP*, volume 2, pages 825–828, 1999.

[11] D. Tran. New Background Modeling for Speaker Verification. In *Proceedings of Interspeech (ICSLP)*, volume 4, pages 2605–2608, 2004.

[12] D. Tran and M. Wagner. A Generalised Normalisation Method for Speaker Verification. In *Proceedings of the Speaker Recognition Workshop (Speaker Odyssey)*, pages 73–76, 2001.

[13] M. Zamalloa, L. J. Rodríguez, M. Peñagarikano, G. Bordel, and J. P. Uribe. Improving robustness in open set speaker identification by shallow source modelling. In *Proceedings of Odyssey 2008: The Speaker and Language Recognition Workshop*, 2008.