Improving robustness in open set speaker identification by shallow source modelling

M. Zamalloa[1,2]*, L.J. Rodríguez*[1]*, M. Peñagarikano*[1]*, G. Bordel*[1]*, J.P. Uribe*[2]

(1) GTTS, Departamento de Electricidad y Electrónica, Universidad del País Vasco
(2) Ikerlan – Centro de Investigaciones Tecnológicas

mzamalloa001@ikasle.ehu.es

Abstract

Open set speaker identification consists of deciding whether an input utterance corresponds to a target speaker or to an impostor. The most likely among a set of target speakers is hypothesized and verified. Speaker verification is performed by comparing the likelihood score of the most likely speaker model to the likelihood score of an impostor model, and then applying a suitable threshold. The most common approach to modelling impostors is the Universal Background Model (UBM). 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 paper, a simple approach is proposed which estimates a shallow source model (SSM) based on the input utterance, and then uses this SSM to normalize the speaker score. Though the SSM does not outperform the UBM, the combination of both models improves the recognition performance and drastically increases the robustness to signals not covered by the UBM.

1. Introduction

Closed-set speaker identification can be easily performed by first training acoustic models for a set of target speakers and then selecting the most likely speaker for each input utterance. But open-set speaker identification involves speaker verification, that is, deciding whether the input utterance was *actually* produced by the most likely speaker or by an impostor. This task may arise in smart non-intrusive environments which must be permanently aware of the potential users, reacting in different ways, with different allowed functionalities, depending on the detected user. If an impostor was detected, the smart environment may automatically block its functionalities or alert the system supervisor. Another interesting application is speaker tracking in broadcast news: the audio signal is segmented into homogeneous sections (usually speaker turns), which must be 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.).

Whatever the application, speaker data are available for a set of target speakers, and speaker models can be trained on them. Though speaker characteristics are reflected at many levels (acoustic, phonetic, phonological, prosodic, syntactic or even pragmatic), and all of them may help the identification task [1], most systems take into account only the physiological information conveyed by the acoustic parameters, and use an acoustic model to gather the statistics of the power spectrum specific to each speaker. Once the acoustic models λ_s are estimated for the set of speakers $s = 1, \ldots, S$, each input utterance X, which consists of a sequence of acoustic vectors $X = \{x_1, x_2, \ldots, x_T\}$, is classified by selecting the most likely speaker $\hat{s}$. Applying the Bayes rule, assuming that all the speakers have equal prior probabilities and the acoustic observations are independent, and taking logarithms, it follows:

$$\begin{aligned} \hat{s} &= \arg \max_{s=1,\ldots,S} P(\lambda_s|X) \\ &= \arg \max_{s=1,\ldots,S} P(X|\lambda_s)P(\lambda_s) \\ &= \arg \max_{s=1,\ldots,S} \log P(X|\lambda_s) \\ &= \arg \max_{s=1,\ldots,S} \sum_{t=1}^{T} \log p(x_t|\lambda_s) \end{aligned} \quad (1)$$

The acoustic *pdf* $p(x|\lambda)$ is usually implemented by a *Gaussian Mixture Model* (GMM) [2]. Once the most likely speaker $\hat{s}$ is determined, verification may be done by comparing the average log-likelihood score $\mathcal{L}(X|\lambda_{\hat{s}}) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t|\lambda_{\hat{s}})$ to a speaker-dependent threshold $\tau(\hat{s})$. The normalizing term $1/T$ is needed to allow applying a length-independent threshold. But the likelihood score not only depends on the speaker but also on many non-speaker utterance-specific variations, so defining a threshold is not a solution, even if we define speaker-dependent thresholds.

To compensate the effect of non-speaker utterance-specific variability and, simultaneously, to allow applying a speaker-independent threshold τ, speaker scores are normalized by the likelihood score of an *impostor model* $\lambda_{\hat{s},I}$:

$$\Lambda(X, \hat{s}) = \mathcal{L}(X|\lambda_{\hat{s}}) - \mathcal{L}(X|\lambda_{\hat{s},I}) \quad (2)$$

In the framework of an open set speaker identification task, the input utterance X is assigned the label $\hat{s}$ if $\Lambda(X, \hat{s}) > \tau$; otherwise, X is taken as an impostor utterance. The decision threshold τ can be heuristically adjusted to trade-off the false acceptance and the false rejection errors. In this case, a false acceptance error corresponds to accepting an impostor as a target speaker, and false rejection errors correspond either to taking a target speaker as an impostor or to taking a target speaker A as the target speaker B (in brief, false rejection errors correspond to missing target speakers).

Various alternatives have been proposed in the literature to define a suitable model for impostors $\lambda_{s,I}$. A possible solution consists of using a cohort of *background speakers* [3]. Background speakers are, in fact, known speakers selected according to a given criterion of closeness, remoteness, competitiveness or the like, with regard to the target speaker. A speaker model is estimated for each background speaker, so that the likelihood

score of impostors is computed as a function (usually the arithmetic mean) of the likelihood scores of background speakers. Two issues arise with this approach: (1) a suitable cohort of background speakers must be selected for each target speaker; and (2) it is not easy to cover all the potential impostors with just a few background speakers.

The most common approach to modelling impostors consists of using a large and diverse (with regard to all the possible sources of variability: gender, age, dialect, etc.) pool of speakers to train a single speaker-independent model, λ_B, called *Universal Background Model* (UBM), usually a GMM with a large number of components [4], 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 speakers; (2) it provides universal acoustic coverage; and (3) it can be used as prior to estimate 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. On the other hand, the verification performance would degrade if impostors, or whatever sources that generate the input signals, were not suitably modelled by the UBM.

If the input signal X was actually generated by the most likely speaker $\hat{s}$, then the likelihood score yielded by $\lambda_{\hat{s}}$ should be much higher than that yielded by λ_B, since $\lambda_{\hat{s}}$ models *specifically one* source, whereas λ_B models *all* the potential sources (both target speakers and impostors). On the other hand, if the input signal X was generated by an impostor, close *but different* to $\hat{s}$, then the likelihood score yielded by $\lambda_{\hat{s}}$ would probably be slightly higher (or even lower) than that yielded by λ_B, because the UBM provides universal acoustic coverage. However, if an impostor utterance was not suitably modelled by the UBM, then the likelihood score of $\lambda_{\hat{s}}$ could still be much higher than that of λ_B, and the utterance could be mistakenly given the label $\hat{s}$.

In this paper we present a new approach to the issue of normalizing speaker scores in speaker verification. Instead of taking as reference an estimation of what input signals should be like (the UBM), we take as reference an estimation of the source based on the input signal. We estimate the acoustic model of the source that generates the input utterance, that we call *Shallow Source Model* (SSM), and then use this SSM to normalize the speaker score, obtaining a measure of how well the speaker model approximates the source model. This approach solves the issue of coverage, since the SSM just attempts to model the source that generates the input utterance. Additionally, it only requires speech data from the target speakers (and obviously, the input signals), but not those additional hours needed to train the UBM. A primitive version of this idea has been successfully applied to speaker tracking in broadcast news [5].

Few alternatives to background speaker models, such as the one presented in this paper, can be found in the literature. It is worth mentioning the work of Hsu, Yu and Yang [6], which is somehow related to our work, since it estimates an acoustic model from the input utterance and takes it as reference to make the decision, but the verification procedure they propose, based on the tolerance interval analysis, use speaker samples instead of speaker models.

The rest of the paper is organized as follows. Section 2 briefly describes the SSM and suggests ways of combining the UBM and the SSM to get a more robust source model. Section 3 gives details about the speaker database, the acoustic parameters and the baseline system used in the experiments. Results are presented and discussed in Section 4, including a test set not modelled by the UBM which reveals the usefulness of the SSM. Finally, conclusions and guidelines for future work are summarized in Section 5.

2. The Shallow Source Model

As explained in the previous section, state-of-the-art speaker verification systems are based on a likelihood ratio, where the likelihood of the claimed speaker is normalized by the likelihood of impostors. Normalizing the speaker likelihood score allows to minimize the effect of non-speaker utterance-specific variability, and a single threshold can be set for all the speakers [4]. However, whereas speaker models are well defined, it is not clear what an impostor model should be and how it could be estimated, since speech data from the *actual* impostors are not available beforehand.

Both cohort models and the UBM aim to model unknown sources, i.e. unknown speakers, by using known data. In particular, if a large and diverse speaker database is used to estimate the UBM, input utterances (either from target or impostor speakers) will be robustly modelled. But it is not always possible to gather enough data to estimate a robust UBM. Also note that, depending on the application (for example, speaker tracking), the source could be non-human (music, noise, etc.). Non-human utterances could be discarded by applying an absolute threshold to the likelihood score. However, in this paper we pursue an alternative for the case an impostor, or whatever source that generates the input utterance, was not suitably modelled by the UBM. In this case neither the speaker model nor the UBM would cover the input utterance and the likelihood ratio would not be reliable.

To improve the robustness to uncovered inputs, instead of modelling all the potential sources by using lots of data, we propose to model just the source that generates the input utterance. We estimate a GMM λ_X from the input utterance X. Since X is usually short (2-10 seconds), a low-order GMM is used 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 kinds of sources). Here we make the assumption that using too many mixture components would model utterance-specific variations instead of source-generic features (we show results in Section 4.1 that support this assumption). In summary, a very simple and shallow GMM, which we call *Shallow Source Model* (SSM), is estimated to model the source.

If λ_X was a *perfect* source model, then it should be:

$$P(X|\lambda_X) > P(X|\lambda_s) \quad \forall s \tag{3}$$

In these conditions, the difference $\Lambda(X,\hat{s}) = \mathcal{L}(X|\lambda_{\hat{s}}) - \mathcal{L}(X|\lambda_X)$ would be always negative or zero, and it would be zero only in the case the speaker model $\lambda_{\hat{s}}$ perfectly matched the source model λ_X. Clearly, in this latter case the speaker $\hat{s}$ should be positively verified, but the same decision should be made if $\Lambda(X,\hat{s})$ was close enough to zero. Using the source model score to normalize the speaker score gives a measure of how well the speaker model approximates the source model. If the log-likelihood ratio $\Lambda(X,\hat{s})$ was greater than a heuristic threshold τ, then X would be assigned the label $\hat{s}$; otherwise, it would be taken as an impostor utterance.

In practice, however, λ_X is not a perfect but a shallow source model and the inequality 3 does not hold. Speaker models are 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 heuristic 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, the SSM would still *guarantee* acoustic coverage to some degree. The likelihood score of the SSM would be higher than that of the most likely 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. Two issues arise, however: (1) the SSM estimates may be highly influenced by utterance-specific variations, so that they would not be robustly modelling the source; and (2) during recognition a new SSM must be estimated for each input utterance, whereas the UBM is estimated beforehand.

To overcome the coverage issue of the UBM and the robustness issue of the SSM, a mixed background model may be estimated by Bayesian adaptation of the UBM to the input utterance. This would take more computation than simply estimating the SSM, since all the parameters of the UBM should be adapted to each input utterance. So, in this work a different approach is proposed, which consists of computing the log-likelihood of impostors as a suitable linear combination of the log-likelihoods of UBM and SSM:

$$\mathcal{L}(X|\lambda_I) = \alpha\mathcal{L}(X|\lambda_B) + (1-\alpha)\mathcal{L}(X|\lambda_X) \qquad (4)$$

where α is a heuristically fixed mixing factor. A somehow similar approach was previously proposed by Tran and Wagner [7], where a constant value $\epsilon > 0$ was added to the likelihood score of the background model, which was shown to reduce false acceptances due to unmodelled inputs.

3. Experimental setup

3.1. Datasets

A phonetically balanced database in Spanish, called Albayzín [8], was used in the experiments. Albayzín, recorded at 16 kHz in laboratory conditions, was originally designed to train acoustic models for speech recognition and is somehow equivalent to TIMIT. Albayzín contains 204 speakers, each speaker contributing at least 25 read utterances and each utterance lasting an average of 3.55 seconds.

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 ones being used as background speakers. Three disjoint sets of utterances were considered: (1) the *training set*, consisting of 15 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 10 utterances from each target speaker and 10 utterances from each impostor, was used to evaluate the performance of the open-set speaker identification systems.

Two different configurations were considered, with 68 and 102 background speakers. The first configuration, *34/68/68*, consists of 510 training utterances, 1700 background utterances and 1020 test utterances (from which 340 correspond to target speakers and 680 to impostors). The second configuration, *34/102/68*, only differs in the background dataset, which consists of 2550 utterances.

Besides Albayzín, a separate database was created to check the robustness to unmodelled inputs. This corpus, called *Mismatched*, is composed of three subcorpora: (1) *Music*, consisting of 288 song fragments taken at random from a song database; (2) *Telephone*, consisting of 340 spontaneous speech fragments taken at random from *Dihana* [9], a database of human-computer dialogues recorded at 8 kHz through telephone lines; and (3) *WWW*, consisting of 332 audio fragments (most of them including speech) taken at random from the internet. So, the whole dataset consists of 960 utterances, all of them lasting 3 seconds, which makes it similar (in size) to the test set of Albayzín.

3.2. Acoustic parameters

Albayzín was originally acquired at 16KHz, so the utterances included in the Mismatched dataset were all resampled at 16 Khz. Each utterance was then analyzed in frames of 25 milliseconds (400 samples), at intervals of 10 milliseconds. A Hamming window was applied and a 512-point FFT computed. The FFT amplitudes were then averaged in 24 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. A Discrete Cosine Transform (DCT) was finally applied to the logarithm of the filter amplitudes, obtaining 12 Mel Frequency Cepstral Coefficients (MFCC). To increase robustness against channel distortion, Cepstral Mean Normalization (CMN) [10] was applied on a utterance by utterance basis. The frame energy was also computed, yielding a 13-dimensional feature vector.

3.3. The baseline system

The baseline system employs the state-of-the-art GMM/UBM paradigm. Gaussian mixture speaker models are directly estimated on the training set. Maximum Likelihood estimates of the GMM parameters are computed using the EM algorithm, starting from random values. Taking into account the size of the training set (15 utterances/speaker, 3.55 seconds/utterance on average), 32-component GMM have been used as speaker models. Regarding the UBM, the optimal size of the GMM has been determined in preliminary open-set speaker recognition experiments (not shown here). For the 34/68/68 and 34/102/68 configurations, the best performance was obtained with 128 and 1024 mixture components, respectively.

3.4. Performance evaluation

To compare the performance of open-set speaker recognition 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 [11], with some modifications that take into account not only speaker verification but also speaker recognition errors, as explained in Section 1. In brief, if an input utterance corresponding to the target speaker A is recognized and verified as corresponding to the target speaker B, then a false rejection error is counted. Sometimes, an optimal operation point in the DET curve is required, such as the EER (*Equal Error Rate*, the point where $P_{\mathbf{miss|target}} = P_{\mathbf{fa|non-target}}$). Here we use the well-known detection cost function used in the NIST speaker recognition evaluations [12]:

$$\begin{aligned} C_{\mathbf{det}} &= C_{\mathbf{miss}} \cdot P_{\mathbf{miss|target}} \cdot P_{\mathbf{target}} + \\ &\quad C_{\mathbf{fa}} \cdot P_{\mathbf{fa|non-target}} \cdot (1 - P_{\mathbf{target}}) \end{aligned} \qquad (5)$$

where $C_{\mathbf{miss}}$ and $C_{\mathbf{fa}}$ are the task-dependent costs of misses

(false rejection errors) and false alarms (false acceptance errors), respectively; $P_{\mathbf{target}}$ is the prior probability of detecting a target speaker; and $P_{\mathbf{miss|target}}$ and $P_{\mathbf{fa|non-target}}$ are experimental values taken from the DET curve. Given the task-dependent costs and the prior probability of detecting a target speaker, the point of the DET curve that minimizes $C_{\mathbf{det}}$ is considered optimal. In the experiments presentd in this paper, $C_{\mathbf{miss}} = C_{\mathbf{fa}} = 1$ and $P_{\mathbf{target}} = 0.33$ (which is the proportion of target speakers in the test set).

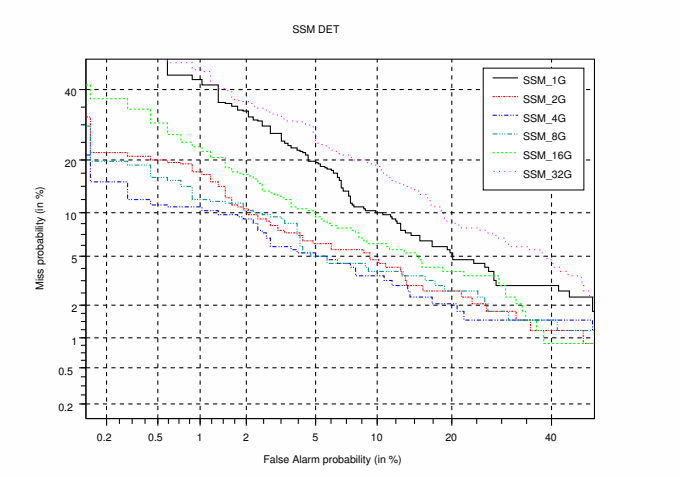

Figure 1: DET curves for six different open-set speaker identification systems using 32-component GMM as speaker models and SSM with 1, 2, 4, 8, 16 and 32 mixture components to normalize speaker scores.

4. Results and discussion

4.1. Tuning the SSM

A first series of experiments was run to determine the optimal size of the GMM used to represent the source in the SSM approach. It is expected to be a low value, since —as was hypothesized in Section 2— a large GMM would be too focused on utterance-specific features. Figure 1 shows DET curves for six different open-set speaker identification systems using 32-component GMM as speaker models and SSM with 1, 2, 4, 8, 16 and 32 mixture components to normalize speaker scores. The best performance was obtained for the SSM with 4 mixture components, which supports our claim for a shallow source model. Note that a 5% EER is obtained without any background information, just the input utterance used to estimate the SSM.

4.2. Improving the UBM with the SSM

Two GMM/UBM systems, UBM1 and UBM2, were developed, corresponding to the configurations 34/68/68 and 34/102/68 described in Section 3.1. Both systems used 32-component GMM as speaker models, and differed in the size of the GMM used as UBM: 128 mixture components (trained on around 6000 seconds of speech) for UBM1, and 1024 mixture components (trained on around 9000 seconds of speech) for UBM2. As shown in Figures 2 and 3, UBM2 clearly outperformed UBM1, which reveals that the performance of UBM is closely related to the acoustic coverage it provides: the larger the training database the better the performance, since a more detailed GMM can be trained and a higher number of potential impostors are covered. Note also that both UBM1 and UBM2 outperformed SSM, though UBM1 was only a little better than SSM. At the point that minimizes the detection cost (marked with circles) UBM1 and SSM show the same miss probability (around 6%), but SSM yields around two times the false alarm probability of UBM1. This suggests that the SSM could be considered as an alternative to the UBM only when not enough data were available to train this latter.

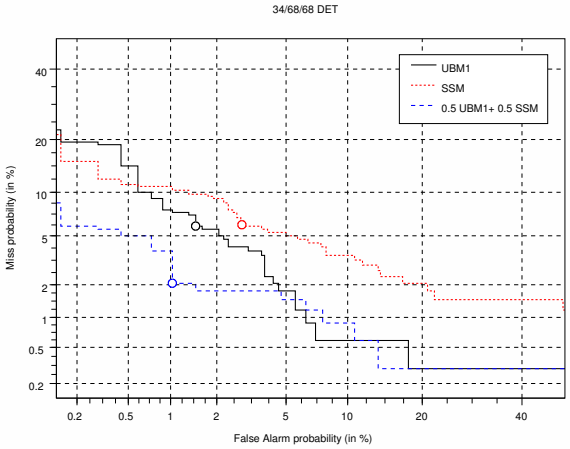

Figure 2: DET curves for three open-set speaker identification systems using 32-component GMM as speaker models and three different likelihood normalization methods: UBM1, a 128-component mixture trained on 68 background speakers; SSM, a 4-component GMM trained on the input utterance; and the optimal combination of the two latter, found for $\alpha = 0.5$.

As proposed in Section 2.1, UBM and SSM could be combined to overcome their respective limitations. UBM1 and UBM2 have been combined with SSM according to Equation 4, for $\alpha = 0.1, 0.2, \ldots, 0.9$, and the performance of the resulting systems has been evaluated. DET curves for the best combinations are shown in Figures 2 and 3. It is worth noting that combining SSM with UBM improves the performance of UBM even in the case of a large 1024-component UBM which clearly outperformed SSM. As could be expected, the improvement was relatively greater for the case of UBM1. As shown in Figure 2, the EER falls from around 5% for SSM and around 4% for UBM1 to less than 2% for the optimal combination of them. The optimal mixing factor can be interpreted as the confidence of the likelihoods provided by the UBM. So, when using a large-mixture UBM, we get $\alpha = 0.9$, whereas for a less robust UBM we get $\alpha = 0.5$.

4.3. Increasing the robustness to unmodelled inputs

Finally, a series of experiments was run to evaluate the robustness of speaker recognition systems to unmodelled inputs. The test set of Albayzín was augmented with the *Mismatched* dataset. The extended test set comprised 1980 utterances, 340 coming from target speakers, 680 from impostors in matched conditions and 960 from impostors in mismatched/unmodelled conditions (see Section 3.1 for details). This way we tried to simulate the situation where a relatively large amount of unmodelled impostor signals must be processed. This is the case of speaker tracking applications, where an input signal is segmented into acoustically homogeneous regions that must be fur-

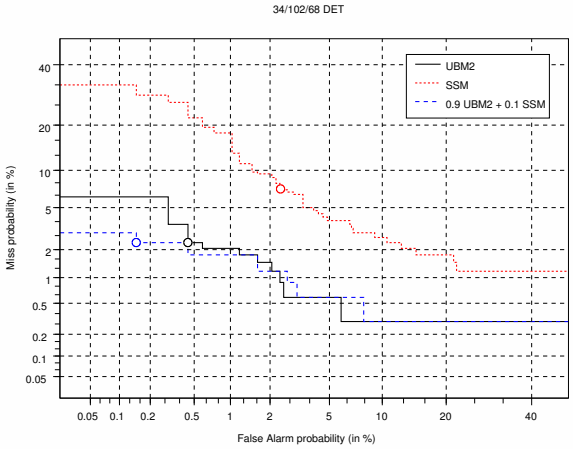

Figure 3: DET curves for three open-set speaker identification systems using 32-component GMM as speaker models and three different likelihood normalization methods: UBM2, a 1024-component mixture trained on 102 background speakers; SSM, a 4-component GMM trained on the input utterance; and the optimal combination of the two latter, found for $\alpha = 0.9$.

ther assigned to a target speaker or to a generic unknown source. Note that only the test set is changed; the speaker and background models are the same used for the experiments in Figure 3 (UBM2, SSM and the optimal combination of UBM2 and SSM).

Figure 4 shows the resulting DET curves. In this case, the circles do not mark the points (i.e. the thresholds) that minimize the detection cost for the extended test set, but those that minimized the detection cost for the original test set (see Table 1). This way, by comparing Figures 3 and 4, the movement of the optimally tuned systems can be followed, providing a meaningful interpretation of the DET curves.

Table 1: Miss and false alarm probabilities and log-likelihood ratio threshold minimizing detection cost for UBM2, SSM and the optimal combination UBM2-SSM (see Figure 3).

	$P_{\mathbf{fa}}$	$P_{\mathbf{miss}}$	Threshold
UBM2	0.0045	0.024	-0.01252
SSM	0.025	0.071	-0.77262
UBM2-SSM	0.0015	0.025	-0.06704

A very important result regards the rejection rate of the unmodelled/mismatched impostor utterances: the SSM system rejects *all of them*, whereas the UBM system accepts 20.5% as target utterances (see Table 2 for details). This makes the DET curve of SSM to improve its false alarm rate with regard to the DET curve shown in Figure 3, while keeping the miss rate, since the 340 target utterances are classified the same way. On the other hand, the false alarm rate of UBM2 degrades so much that the DET curve moves rightwards in around 15 absolute points. As a result, SSM clearly outperforms UBM2 when dealing with the extended test set.

However, the most relevant result from Figure 4 is that the combination of UBM2 and SSM further improves the performance of SSM, yielding the best DET curve. This means that, though the performance of UBM2 was strongly degraded when dealing with unmodelled/mismatched impostor utterances, the

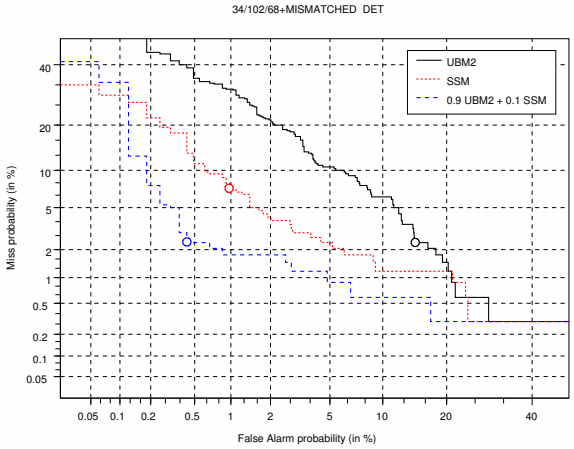

Figure 4: DET curves for three open-set speaker identification systems (UBM2, SSM and the optimal combination of UBM2 and SSM), computed over an extended test set which includes signals not covered by the UBM. Circles mark the optimal thresholds for the original test set.

Table 2: False alarm probabilities over the whole *Mismatched* dataset and the subsets *Music*, *Telephone* and *WWW*, computed at the points minimizing detection cost for the original test set (see Table 1), for UBM2, SSM and the optimal combination UBM2-SSM.

	$P_{\mathbf{fa}}$			
	Mismatched	Music	Telephone	WWW
UBM2	0.205	0.475	0.047	0.130
SSM	0.000	0.000	0.000	0.000
UBM2-SSM	0.003	0.000	0.006	0.003

SSM helped reject them and the performance of the combination suffered little degradation with regard to that of Figure 3. Table 2 shows the false alarm probabilities for the whole *Mismatched* dataset and for the three subsets *Music*, *Telephone* and *WWW*. In the case of UBM2, the highest false alarm probability was found for the *Music* subset (0.475), whereas the *Telephone* subset (consisting of speech through telephone channel) yielded the smallest value (0.047). On the other hand, the SSM approach yielded zero false alarm probability for all the subsets. In any case, when balancing results for both modelled and unmodelled inputs, the combination UBM-SSM provided the best speaker recognition performance.

5. Conclusions

Using a background model to normalize speaker likelihood scores is a common practice in speaker verification. Most systems use a large pool of speaker data to estimate a single background model, called *Universal Background Model* (UBM), usually a GMM with a large number of components, which is used to normalize the likelihoods of all the target speakers. The UBM attempts to cover the acoustics of all the potential impostors. However, it is not always possible to gather enough data to estimate such a robust UBM. On the other hand, the verification performance would degrade if the input signals were not suitably modelled by the UBM.

In this paper we have presented a new approach to the issue of normalizing speaker scores in speaker verification which aims to improve the robustness to uncovered sources. Instead of modelling all the potential sources by using lots of data, we estimate a low-order GMM from the input utterance, which we call *Shallow Source Model* (SSM), and then use this SSM to normalize the speaker score. This approach solves the issue of coverage, since it just attempts to model the source that generates the input utterance. Additionally, it only requires speech data from the target speakers (and obviously, the input signals), but not those additional hours needed to train the UBM.

Open-set speaker recognition experiments have been presented which evaluate the performance of SSM and compare it to UBM. Though SSM did not outperform UBM, it was found that a suitable combination of the SSM and the UBM likelihoods improved the performance of UBM even in the case of a large UBM. Finally, a series of experiments was run to evaluate the robustness of speaker recognition systems to unmodelled inputs. In this case, SSM clearly outperformed UBM, but the most relevant result was that the combination of UBM and SSM further improved the performance of SSM. This means that, though the performance of UBM was strongly degraded when dealing with unmodelled impostor utterances, the SSM helped reject them and the performance of the combination suffered a little degradation. In any case, the combination UBM-SSM provided the best speaker recognition performance either with modelled or with unmodelled signals.

Future work includes comparing the weighted combination of UBM and SSM to the estimation of a single source model by Bayesian adaptation of the UBM to the input utterance. Also, more exhaustive experimentation is planned by rotating target speakers, background speakers and impostors. Finally, the SSM approach will be tested in more realistic conditions, by using speaker databases recorded over telephone channels.

6. Acknowledgements

This work has been partially funded by the Government of the Basque Country, under program SAIOTEK, projects S-PE05IK06, S-PE05UN32 and S-PE06UN48, and the University of the Basque Country, under project EHU06/96.

7. References

[1] Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., and Xiang, B., "The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition", Proceedings of ICASSP 2003, Vol. IV, pp. 784–787.

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

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

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

[5] Rodríguez, L. J., Peñagarikano, M., and Bordel, G., "A Simple But Effective Approach to Speaker Tracking in Broadcast News", J. Martí et al. (Eds.): IbPRIA 2007, Part II, LNCS 4478, pp. 48–55.

[6] Hsu, C.-N., Yu, H.-C., and Yang, B.-H., "Speaker Verification Without Background Speaker Models", Proceedings of ICASSP 2003, Vol. II, pp. 233–236.

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

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

[9] Alcocer, N., Castro, M. J., Galiano, I., Granel, R., Grau, S., and Griol D., "Adquisición de un Corpus de Diálogo: DIHANA", III Jornadas en Tecnología del Habla, pp. 131–134, Valencia (Spain), November 2004.

[10] Rosenberg, A. E., Lee, C.-H., and Soong, F. K., "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", Proceedings of the ICSLP'94, pp. 1835–1838, Yokohama, Japan, 1994.

[11] http://www.nist.gov/speech/tools/DETware_v2.1.targz.htm

[12] Doddington, G. R., Przybocki, M. A., Martin, A. F., and Reynolds, D. A., "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective" Speech Communication, 31(2-3):225–254, June 2000.