A model space framework for efficient speaker detection

Mathieu Ben, Guillaume Gravier, Frédéric Bimbot

Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)
CNRS and INRIA
Equipe Modélisation et Expérimentation pour le Traitement de l'Information et du Signal Sonores
mathieu.ben,frederic.bimbot,guillaume.gravier@irisa.fr

Abstract

In this paper, we investigate the use of a distance between Gaussian mixture models for speaker detection. The proposed distance is derived from the KL divergences and is defined as an Euclidean distance in a particular model space. This distance is simply computable directly from the model parameters thus leading to a very efficient scoring process. This new framework for scoring is compared to the classical log likelihood ratio score approach on a speaker verification task of the NIST 2004 evaluation and on the speaker tracking task of the ESTER french evaluation. Results shows that the proposed approach is competitive and leads to computation times divided by a factor of more than 3.

1. Introduction

Speaker detection consists in deciding if a given speaker is present in a test material. If the test material is supposed to be from a single speaker, the task is called speaker verification. Speaker detection can be associated to an audio segmentation process in a speaker tracking system. In that case, the task consists in detecting which of the target speakers are present in each audio segment.

Currently, most of the speaker detection systems rely on the UBM-GMM approach [1], which is now sometimes associated to other classifiers like support vector machines. In the UBM-GMM framework, target speaker Gaussian mixture models (GMM) are derived from a universal background model (UBM) with *maximum a posteriori* (MAP) adaptation, usually limited to mean re-estimation. Given a training utterance, prior and posterior means for each Gaussian in the mixture are linearly combined to get the MAP estimated means. The balance between prior and posterior parameters is controlled by the occupation rate of each Gaussian and by a relevance factor τ which controls the adaptation. This maximization problem is iteratively achieved via the EM algorithm using the following re-estimation formulae for mean m_k in Gaussian k :

$$m_k = \frac{\gamma_k}{\gamma_k + \tau}\overline{y_k} + \frac{\tau}{\gamma_k + \tau}\mu_k \quad . \tag{1}$$

In this equation, γ_k is the occupation rate of Gaussian k, and $\overline{y_k}$ and μ_k are respectively the posterior and prior mean for Gaussian k.

The decision stage is based on a log-likelihood ratio (LLR) detector in which the UBM is used as a common impostor model $\hat{P}_\Omega$ for all target speakers. Given a test material $\mathcal{Y}$ (an utterance or a speech segment) and a target model $\hat{P}_X$, the detection score $S_X(\mathcal{Y})$ is compared to a fixed threshold, eventually after applying a score normalization. If the score is greater than the threshold θ the target speaker is detected, either it is assumed not present in the test material :

$$S_X(\mathcal{Y}) = \log \frac{\hat{P}_X(\mathcal{Y})}{\hat{P}_\Omega(\mathcal{Y})} \underset{reject}{\overset{detect}{\gtrless}} \theta \quad . \tag{2}$$

This approach has shown good performance, as illustrated in the NIST speaker recognition evaluations (SRE)[1], in particular when normalization techniques are applied at the feature and score levels to improve system robustness. However, likelihood computations are costly and may lead to unacceptable testing time if the number of target speaker is large or in the case of a low capacity hardware architecture. Moreover, "online" score normalization like T-norm [2] significantly increase computation time during test, even when the N-best technique [1] is used to estimate LLRs.

To allow fast and "light" scoring, we investigate the use of a simple distance between models to compute detection scores, instead of the classical LLRs. We propose to use an Euclidean distance between mean-adapted GMMs, which is derived from the Kullback-Leibler divergence, and which only depends on the model parameters.

We motivate our choice in section 2 where we first show that LLR scores are linked to KL divergences. We then define a similarity measure between speaker models using an upper bound of the KL divergence in the case of mean only adapted GMMs. This similarity measure, which is homogeneous to a squared Euclidean distance in a particular model space, can be used to efficiently compute scores, thus defining a new framework for speaker detection. Experimental application of this model space framework is given in section 3 in the context of the NIST 2004 SRE and in the ESTER [5] phase 2 speaker tracking task. Results show that the new framework for score computation is competitive with the classical LLR approach and significantly decrease computation time during test. The conclusion is given in section 4 where we also point the numerous perspectives of this work.

2. Model space speaker characterization

2.1. Motivations : link between LLRs and KL divergences

Classically, the average frame LLR over the whole test material $\mathcal{Y} = \{y_1, ..., y_n\}$ is used as a raw score for speaker detection :

$$S_X(\mathcal{Y}) = \frac{1}{N}\sum_{n=1}^{N} \log \frac{\hat{P}_X(y_n)}{\hat{P}_\Omega(y_n)} \quad . \tag{3}$$

[1] http://www.nist.gov/speech/tests/spk

Introducing the "true" model $\mathcal{M}_Y$ of the test class, this score is linked to Kullback-Leibler divergences as follows :

$$E_{\mathcal{M}_Y}[S_X(\mathcal{Y})] = E_{\mathcal{M}_Y}[\log \frac{\mathcal{M}_Y(y)}{\hat{P}_\Omega(y)}] + E_{\mathcal{M}_Y}[\log \frac{\hat{P}_X(y)}{\mathcal{M}_Y(y)}] \ , \\ = KL[\mathcal{M}_Y||\hat{P}_\Omega] - KL[\mathcal{M}_Y||\hat{P}_X] \ , \quad (4)$$

where $E_{\mathcal{M}_Y}[.]$ denotes the expectation calculated under model $\mathcal{M}_Y$.

In a detection task context, $\mathcal{M}_Y$ is either the "true" target model $\mathcal{M}_X$, or the "true" impostor model $\mathcal{M}_{\bar{X}}$. In the case of client accesses (i.e. $\mathcal{M}_Y = \mathcal{M}_X$), the first divergence in the above equation relates to the discriminative power of the (true) client class with respect to the world model. The second divergence is a penalizing term which pull client scores towards negative values. It relates to the quality of the estimated client model $\hat{P}_X(y)$. Considering now impostor accesses (i.e. $\mathcal{M}_Y = \mathcal{M}_{\bar{X}}$), the first divergence is related to the quality of the estimated impostor model, represented by the world model $\hat{P}_\Omega(y)$, and it penalizes the impostor scores towards positive values. The second divergence relates to the discriminative power of the (true) impostor class with respect to the estimated client model.

This interpretation of the expected verification scores shows that they are asymptotically equivalent to differences between two KL divergences. It suggests that verification scores could be calculated from similarity measures between models, instead of the classical LLR scores. In a task involving a large number of score calculations for each test, for example in the case of speaker identification, tracking or verification with T-norm scores, a simply computable distance between models should lead to an important gain in testing time. However, in the case of GMMs, KL divergences must be estimated with a Monte-Carlo method, which is computationally expensive. In the next section we show how a simple Euclidean distance derived from the KL divergence can be used instead, leading to the definition of a new framework for speaker detection in a model space.

2.2. An Euclidean distance between mean-adapted GMMs

Considering mean-only adapted GMMs, it can be shown (see [3] and [4]) that the KL symmetric divergence $KL2$ between the two models P_{X_1} and P_{X_2} is upper bounded by

$$KL2(P_{X_1}, P_{X_2}) \leq \sum_{k,d} w_k . \frac{(m_{k,d}^{X_1} - \tilde{m}_{k,d}^{X_2})^2}{\sigma_{k,d}^2} \quad (5)$$

where $m_{k,d}^{X_1}$ and $m_{k,d}^{X_2}$ are the d components of the mean vectors of Gaussian k in P_{X_1} and P_{X_2} respectively, w_k is the weight of Gaussian k in both models and $\sigma_{k,d}^2$ is the $d \times d$ elements of their common covariance matrix in Gaussian k. The term on the right hand side of this inequality is a similarity measure between two mean-only adapted GMMs, which is simply defined by the model parameters and which is homogeneous to a KL divergence. Moreover, defining the space of parameters

$$\{\lambda_{k,d}\} = \{\sqrt{w_k}\frac{\Delta m_{k,d}^X}{\sigma_{k,d}}\},$$

where $\Delta m_{k,d}^X = m_{k,d}^X - m_{k,d}^\Omega$, this term is a squared Euclidean distance between the two points representing models P_{X_1} and P_{X_2}. We will denote $D_E(P_{X_1}, P_{X_2})^2$ this squared Euclidean distance. In the space defined by the $\{\lambda_{k,d}\}$ parameters, the world model corresponds to the origin (see figure 1).

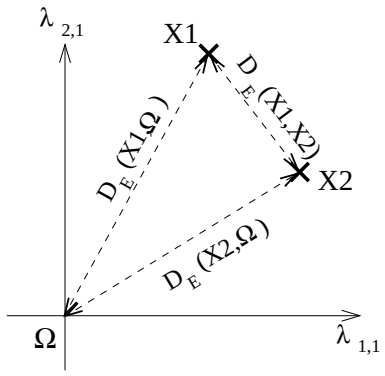

Figure 1: Representations of GMMs in the model space with the corresponding Euclidean distances

One advantage of this Euclidean distance is that it can be computed directly from the parameters of the GMMs, with a low computational cost (no log or exponential calculation). Moreover it is highly related to KL divergences according to relation (5). In practice, it has been observed that ${D_E}^2$ and $KL2$ are strongly correlated (with a correlation coefficient of 0.99 obtained with the speaker set of NIST'04 SRE), indicating that they bring equivalent information about model similarity.

In the next section we show how we use the ${D_E}^2$ metric to compute detection scores.

2.3. Detection scores in the model space

In a similar way as in equation (4), we define a detection score as the difference between two similarity measures, replacing the KL divergences by the ${D_E}^2$ metric and using a model $\hat{P}_Y$ estimated on the test material :

$$S_X^{(D_E)} = D_E(\hat{P}_Y, \hat{P}_\Omega)^2 - D_E(\hat{P}_Y, \hat{P}_X)^2 \quad (6)$$

This score can be efficiently computed and is linked to the LLR score via relations (5) and (4). Note however that it is necessary to train a model on the test material before being able to compute a score in the model space, which could seem not advantageous with respect to an LLR computation. But if a large number N of models has to be compared to the test material (which can be the case for speaker tracking or when using T-norm), one has only to first train the test model and then compute N scores as given in equation (6). This should be much more efficient than computing N LLR scores, even when the N-best technique is used.

In section 3 we experimentally validate this model space framework on the NIST 2004 speaker verification task, and in the speaker tracking task of the ESTER phase 2 evaluation.

3. Application to speaker verification and tracking

3.1. Experiments on the NIST 2004 SRE

3.1.1. Evaluation conditions

We present here experimental application of the model space framework on the 1side/1side speaker verification task of the

NIST 2004 SRE. In this task, the speech of a single speaker and extracted from a 5 minute conversation is available for training and for testing. The data is part of the FISHER corpus collected by the LDC. It is mainly composed of telephone conversations in English language but a small part of the tests are concerned with other languages. Landline and cellular phones are used in this data set.

3.1.2. System descriptions

Baseline system
The baseline system is a classical GMM-UBM system as described in the introduction.

Feature vectors are composed of 16 cepstral coefficients along with their 16 Δ coefficients and the $\Delta -$ log energy. A frame selection process is applied based on a bi-Gaussian modeling of the frame energy distribution and an ML classification of each frame into speech or silence class. The remaining feature vectors are normalized to a zero-mean, unit-variance distribution on a 3 second sliding window.

Two gender dependent background models are used, estimated using about 4h30 of speech for each one, from the NIST'01 and NIST'02 corpora (mixed cellular and landline data). These models are 256 component GMMs with diagonal covariance matrices. Target speaker models are adapted from the background models using a MAP criterion. Only the mean vectors are adapted with a single EM iteration.

The raw verification scores are computed using the 10 best Gaussians in the background model and then normalized with T-norm using impostor models from the NIST'01 and NIST'02 training sets (50 male and 50 female speakers with cellular data, plus 50 male and 50 female speakers with landline data).

Model space system
The model space (MS) system mainly differs from the baseline by its scoring procedure which is described in section 2. Models of the test utterances are estimated with the same procedure as for the target speaker models (mean-only MAP adaptation with one iteration of EM), and verification scores are evaluated throw the computation of the squared Euclidean distances ${D_E}^2$ (see equation 6). However, in practice, it has appeared that a model normalization process is necessary to make the MS system work. This is due to the fact that, as speaker and test models are adapted from a background model Ω, the distance between a given model X and the origin in the model space (which corresponds to Ω) highly depends on the amount of data used to train the model X. As a consequence, scores in the model space may be very heterogeneous and lead to poor performances.

To make the scores more homogeneous we apply a very simple normalization in the model space, that we called M-norm ("Model normalization"). It consists in projecting points representing the models on a unit hyper-sphere (see figure 2). After normalization, all models are at the same distance D_E (here $D_E = 1$) from the background model. This model normalization relies on the hypothesis that the main speaker specific information is given by the direction that takes a model in the model space and that the original distance between the model and the origin is a pertubating information for the reasons given above. The detection scores are then computed with normalized target speaker models according to equation (6) and the T-norm scores are computed with normalized impostor models. Note that this model normalization does not require any additional normalization data.

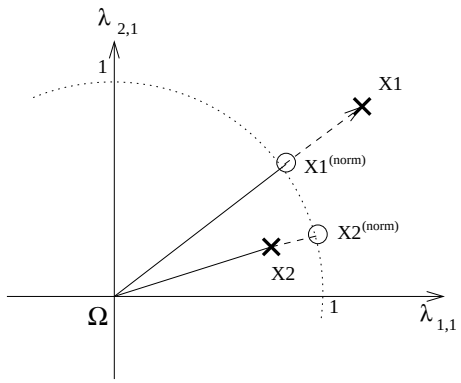

Figure 2: Illustration of the model normalization procedure

3.1.3. Results

The results obtained by the baseline LLR system and the MS system with T-norm are illustrated by the detection error trade-off (DET) curves on figure 3, representing the evolution of the false alarm and miss rates of the systems when the decision threshold varies. As mentioned previously, a large performance lost is observed for the MS system when no model normalization is applied, even if the scores are T-normalized. However when the M-norm is used in addition to T-norm, the MS system gives performances as good as the baseline system. A slight gain in performance is observed for the MS system in the low miss rate region. These results validates the proposed approach for scoring in a model space. Moreover computation time during testing[2] was reduced by about 75% when using the MS system, with respect to the baseline LLR/N-best system. This shows that the model space framework proposed in section 2 leads to a very efficient scoring process for speaker verification with T-norm, without lost of performance.

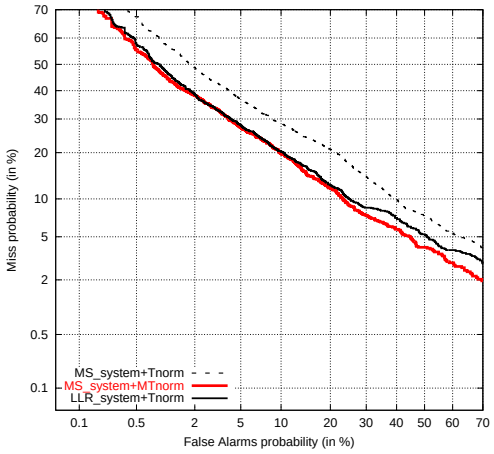

Figure 3: DET curves : performance of the MS system with T-norm and with M-norm+T-norm (MT-norm), and performance of the baseline LLR system with T-norm

[2] all experiments were performed on a Pentium 4 processor at 2.4 GHz (4700 bogomips), with 1Gb of RAM and linux OS

3.2. Experiments on the ESTER speaker tracking task

In this section we present results obtained on a speaker tracking task in the framework of the French speaking evaluation campaign of broadcast news rich transcription systems ESTER [5].

3.3. Task conditions and systems

The data used in the the ESTER phase 2 evaluation consists of broadcast news documents in French language from several radios (refer to [5] for details about the corpus).

For the tracking task, a list of 279 target speakers with at least 2 minutes of speech in the training set was provided. The amount of training speech available for a target speaker varies from 2 minutes to more than an hour. For some speakers, information on the radio(s) in which they appear was also provided. For the evaluation campaign, performance were measured using the F-measure. In this paper, we will report performances using DET curves.

In our system, speaker tracking is achieved in two sequential steps. An automatic segmentation of the audio stream is first performed, using a speech detector followed by a BIC (Bayesian Information Criterion) based algorithm to detect speaker changes. Then, speaker detection is performed individually for each target speaker referenced in the radio (or with no radio specified), on each identified speech segment. The speaker detection baseline system is derived from the speaker verification system described in section 3.1.2, with specific tunings on the ESTER corpus. The UBM is a gender-independent 512 component diagonal GMM estimated using 13 hours from the training corpus. The detection scores were T-normalized using a cohort of 250 impostor models (117 female speakers and 133 male speakers) taken from the training corpus. Decision thresholds were optimized on the development set so as to maximize the F-measure.

3.4. Results

The DET curves obtained by the speaker tracking systems based on model space (MS) detection scores and LLR detection scores (with T-norm) are plotted on figure 4, for the development (dev) and evaluation (eva) data sets of the ESTER phase 2 evaluation.

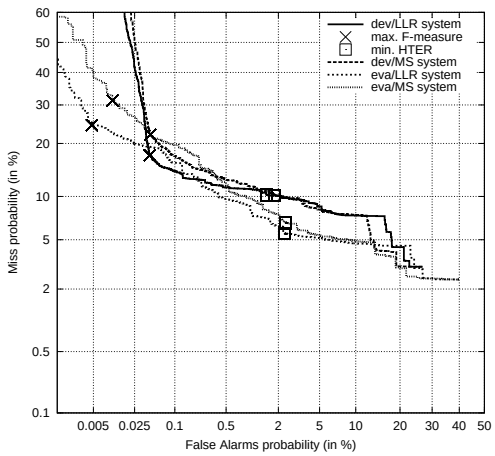

Figure 4: DET curves : performance of the tracking systems based on MS scores and LLR scores (with T-norm).

In these experiments, a slight performance lost is observed for the MS system at the max F-measure point, with respect to the LLR system, in particular on the evaluation set. Additional analysis have shown that this could be due to a bad behavior of the MS system on very short segments. This will be analysed in more details in a future work. No significant differences appears between the two systems at the HTER (Half Total Error Rate) point. Using the MS system for speaker tracking in these experiments resulted in a gain of 60% in testing time, showing here again that this scoring approach is very effective when the test material must be compared to a large number of models.

4. Conclusion and perspectives

We have developed a new framework for efficient scoring in a model space, in a speaker detection context. The proposed approach is based on an Euclidean distance between adapted GMMs which is simply computable from the model parameters. Applied on the NIST 2004 speaker verification task, this model space scoring resulted in reducing computation time by a factor of 4 during test, compared to the classical log-likelihood ratio scoring, without lost of performance. On the ESTER speaker tracking task it resulted in a reduction of 60% of the testing time but a slight degradation was observed at the max F-measure point. However no performance lost appeared at the HTER point in this experiment.

This model space framework offers numerous interesting perspectives. First, the approach allows fast scoring with good performance and could be used in extensive speaker recognition tasks like speaker identification on a very large set or speaker indexing of huge audio databases. Second, the scoring process is relatively light in itself, if we don't consider test model estimation. It could be implemented on embedded hardware architecture like smart card. In this case, the test model estimation step should be performed on a more powerful host terminal.

We also plan to develop compensation techniques in the model space, like channel or handset normalizations. One possible way would be to learn transformations from channel dependent models to a channel independent model. An other more elegant procedure could consist in identifying perturbed directions in the model space, using data analysis techniques like PCA, and to directly eliminate them at the scoring level.

5. References

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, 2000.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for test-independent speaker verification systems," *Digital Signal Processing Vol 10,num 1-3*, 2000.

[3] M. N. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models," in *IEEE Signal Processing Letters*, April 2003.

[4] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms," in *Intl. Conf. on Speech and Language Processing*, 2004.

[5] G. Gravier, J. F. Bonastre, S. Galliano, E. Geoffrois, K. M. Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of french broadcast news," in *Language Evaluation and Resources Conference*, 2004.