AN ALTERNATIVE TO SILENCE REMOVAL FOR TEXT-INDEPENDENT SPEAKER VERIFICATION

Johnny Mariéthoz *Samy Bengio*

IDIAP
CP 592, rue du Simplon 4,
1920 Martigny, Switzerland
{bengio,marietho}@idiap.ch

ABSTRACT

State-of-the-art text independent speaker verification systems use silence/speech detectors to get rid of silence frames which are considered to be non discriminative. This paper explores a possible replacement to this silence/speech detector by considering each Gaussian of a GMM as modeling a specific speech class and by using discriminant models like SVMs and MLPs in order to fuse the corresponding class-specific scores to obtain a final decision. Experiments on the NIST 2001 database yielded statistically significantly better performance for the new model as compared to our best baseline system involving a silence/speech detector, without having to rely on uncertain hypotheses.

1. INTRODUCTION

The goal of speaker verification is to decide whether a given speech utterance was pronounced by a claimed client or by an impostor. One of the steps involved in the classical approach converts the continuous speech signal into a sequence of vectors called frames. These frames can contain silence or speech information but the silence frames are believed not to contain any speaker information. Hence, state-of-the-art text-independent speaker verification systems based on Gaussian Mixture Models (GMMs) normally use a silence/speech detector to discard these frames.

However, using such a detector (which is prone to error since there is no precise definition of a silence frame) as a preprocessing step could remove frames that would have in fact contained useful information with regard to the final task, or could keep some unwanted frames, which, as we will see later in section 2.2, could harm the quality of the resulting model.

Instead of using a silence/speech detector, this paper proposes a new approach: given the hypothesis that each Gaussian of a GMM represents a certain class of speech sub-unit, a discriminant model could be used to weight the scores given by each of these Gaussian/class of speech sub-unit.

The experiments performed on the NIST 2001 database show that the new approach yields results better than the baseline system with the advantage that no a priori knowledge is used regarding the discrimination of the input frames.

The next section describes the baseline system, including the GMM model and the silence/speech detector. The following section exposes the new model based on class-specific Gaussians. Finally some experiments on the 2001 NIST database are presented and analyzed.

2. BASELINE SPEAKER VERIFICATION SYSTEM

Classical speaker verification models are based on a statistical framework. We are interested in $P(S_i|\mathbf{X})$ the probability that a speaker S_i has pronounced sentence $\mathbf{X}$. Using Bayes theorem, we can write it as follows:

$$P(S_i|\mathbf{X}) = \frac{p(\mathbf{X}|S_i)P(S_i)}{p(\mathbf{X})}. \tag{1}$$

In order to decide whether or not S_i has pronounced $\mathbf{X}$, we compare $P(S_i|\mathbf{X})$ to the probability that any other speaker has pronounced $\mathbf{X}$, which we write $P(\bar{S}_i|\mathbf{X})$. When $P(\bar{S}_i|\mathbf{X})$ is the same for all clients, which is the case in this paper, we replace it by a speaker independent model $P(\Omega|\mathbf{X})$ where Ω represents the *world* of all the speakers. The decision rule is then:

$$\text{if } P(S_i|\mathbf{X}) > P(\Omega|\mathbf{X}) \text{ then } \mathbf{X} \text{ was uttered by } S_i. \tag{2}$$

Using equation (1), inequality (2) can then be rewritten as:

$$\frac{p(\mathbf{X}|S_i)}{p(\mathbf{X}|\Omega)} > \frac{P(\Omega)}{P(S_i)} = \delta_i \tag{3}$$

where the ratio of the prior probabilities is usually replaced by a threshold δ_i since it does not depend on $\mathbf{X}$. Taking the logarithm of (3) leads to the *log likelihood ratio*:

$$\text{llr} = \log p(\mathbf{X}|S_i) - \log p(\mathbf{X}|\Omega) > \log \delta_i = \Delta_i. \tag{4}$$

In order to implement this, we usually train a *world model* $p(\mathbf{X}|\Omega)$ by Maximum Likelihood (ML), and then adapt it using a Maximum A Posteriori (MAP) technique for each client model $p(\mathbf{X}|S_i)$ [1]. As the number of client data is limited, we often estimate a client independent decision threshold Δ.

2.1. Gaussian Mixture Models

In the context of text-independent speaker verification, both $p(\mathbf{X}|S_i)$ and $p(\mathbf{X}|\Omega)$ are most often modeled by Gaussian Mixture Models (GMMs) with diagonal covariance matrices. In order to use such a model, we make the assumption that the frames of the speech utterance are independent from each other: the probability of a sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ given a GMM with N Gaussians is computed as follows:

$$p(\mathbf{X}) = \prod_{t=1}^{T} p(\mathbf{x}_t) = \prod_{t=1}^{T} \sum_{n=1}^{N} w_n \; p(\mathbf{x}_t|n) \tag{5}$$

where w_n is the weight of Gaussian n and $p(\mathbf{x}_t|n)$ is the corresponding Gaussian probability density function.

In equation (5), the magnitude of $p(\mathbf{X})$ depends of the number of frames T. In order to use a threshold Δ independent of T, a normalization factor is thus added. Using equations (5) and (4), and adding the normalization factor we obtain:

$$\begin{aligned} \mathrm{llr} &= \frac{1}{T}\sum_{t=1}^{T}\log\sum_{n=1}^{N} w_n\ p(\mathbf{x}_t|S_i,n) \\ &\quad -\frac{1}{T}\sum_{t=1}^{T}\log\sum_{n=1}^{N} w_n\ p(\mathbf{x}_t|\Omega,n) \end{aligned} \tag{6}$$

where $p(\mathbf{x}_t|S_i,n)$ is the density of Gaussian n of the client model and $p(\mathbf{x}_t|\Omega,n)$ is the density of Gaussian n of the world model.

2.2. Silence Frames

In any given sentence, silence often appears between words. These silence segments obviously do not contain much speaker information. Hence, state-of-the-art systems usually remove them with the help of a silence/speech detector. In fact, the main reason to get rid of them is that they influence the normalization factor described in section (2.1): The more there are silence frames, the smaller will be the score after normalization, which is not a desirable behavior of the system, which should be robust to silences.

The simplest approach to remove silences compares the energy of each frame to a given threshold learned on the first few frames [2]. Our own baseline approach learns in an unsupervised way a bi-Gaussian model with the hypothesis that the distribution of the silence parts should be different from the speech part and that the energy of the speech is bigger than the silence one. Afterward, we remove all frames $\mathbf{x}_t$ such that

$$p(\mathbf{x}_t|G_{speech}) < p(\mathbf{x}_t|G_{silence}) \tag{7}$$

where G_{speech} is the Gaussian representing the speech while $G_{silence}$ represents the silence. Note that the bi-Gaussian model is a simple GMM with two Gaussians learned on the complete feature vector. The enrollment is unsupervised because the silence/speech segmentation is unknown. Moreover, while in the simple energy based approach, one needs to use the first few frames (which are hypothesized to be silence) to estimate the level of silence energy, this unsupervised technique does not need such hypothesis. A similar approach is described in details in [2].

3. GAUSSIANS REPRESENT CLASSES

Based on this silence removal approach, we would like to propose the following idea: if a bi-Gaussian model can discriminate between silence and speech frames, the world model (a GMM with more that 100 Gaussians in general) can probably also perform something similar. It is often conjectured that each Gaussian of the world model represents a different class of the signal (silence, noise, vowels, sub-phonemes, etc...) and it is reasonable to think that some classes contain more discriminant information than others. Moreover, removing silence frames can be seen as a hard case where the corresponding soft case would attribute a weight for each class.

For a given client, let us now consider, for each Gaussian i of the world model, the couple containing Gaussian i and its associated Gaussian from the client model. This couple contains discriminative information between client and world regarding a given class.

We already know, that the silence class does not contain discriminant informations. In a similar way, the other classes could have different levels of discrimination.

3.1. From GMMs to Gaussian Scores

In order to consider couples of Gaussians, we first need to enforce an exact correspondence between the world and client Gaussians. This is in fact already the case when using MAP adaptation to train client models.

Let us now assign each frame $\mathbf{x}_t$ to only one Gaussian using a kind of Viterbi approximation of equation (6) as follows. Let $n^*_{t,\theta}$ be the Gaussian in model θ that best represents $\mathbf{x}_t$:

$$n^*_{t,\theta} \quad = \quad \arg\max_n \left[\log w_n\ p(\mathbf{x}_t|\theta,n)\right] \ . \tag{8}$$

We can compute a Viterbi approximation of llr as follows:

$$\mathrm{llr} \approx \mathrm{llr}_v = \frac{1}{T}\sum_t \left[\log p(x_t|S_i,n^*_{t,i}) - \log p(x_t|\Omega,n^*_{t,\Omega})\right] \ . \tag{9}$$

Note that there are no constraint in (9) that guarantees that a given frame is assigned to the same Gaussian in the client and world models. In order to enforce this, a synchronous alignment procedure is used [3]:

$$n^*_t = \arg\max_n \left[\begin{array}{l} \alpha\log w_n\ p(\mathbf{x}_t|S_i,n)+ \\ (1-\alpha)\log w_n\ p(\mathbf{x}_t|\Omega,n) \end{array}\right] \tag{10}$$

where α is a trade-off between placing our confidence in the world or the client model. Using this synchronous alignment, we define a new score as follows:

$$\mathrm{llr}_v \approx \mathrm{llr}_s = \frac{1}{T}\sum_t \left[\log p(\mathbf{x}_t|S_i,n^*_t) - \log p(\mathbf{x}_t|\Omega,n^*_t)\right] \ . \tag{11}$$

We can also compute separately llr_s for each Gaussian n, as follows:

$$\mathrm{llr}_s(n) = \frac{1}{T(n)}\sum_{t:n=n^*_t}\left[\log p(\mathbf{x}_t|S_i,n) - \log p(\mathbf{x}_t|\Omega,n)\right] \tag{12}$$

where $T(n)$ is the number of frames captured by Gaussian n in equation (10). For each access having a variable length, we thus obtain a vector which size is equal to the number of Gaussians (N). Each value of this vector is the score (12) given by the corresponding couple of Gaussians. Hence, this model produces a fixed length vector from a sequence of variable number of frames.

3.2. Fusion of Scores

The goal of this last step is to obtain a single decision score out of the vector of scores obtained using equation (12). The simplest way to merge several scores is to average them. However, in some cases, a more complex fusion process might give better results. For instance, one could use a weighted sum, where the weights would be trained by maximizing the likelihood on a separate training set. Another solution, favored in this paper, is to use powerful machine learning techniques such as Support Vector Machines (SVMs) [4] or Multi-layer Perceptrons (MLPs) [5].

The output score S of the fusion model is then:

$$S = f(\mathrm{llr}_s(1), \ldots, \mathrm{llr}_s(N)|\Theta) \quad (13)$$

where $\mathrm{llr}_s(i)$ is the Gaussian specific log likelihood ratio obtained by equation (12) for Gaussian i, $f(\cdot;\Theta)$ is a set of functions (such as the set of linear functions, or the set of functions induced by an MLP), and Θ is the parameter vector that optimizes a given criterion (such as Equal Error Rate, EER, see experimental section) on a separate development set.

4. EXPERIMENTS

4.1. Database and Protocol

We tested the new speaker verification model on a subset of the database that was used for the *2001 NIST Speaker Recognition Evaluation*, which comes from the Switchboard-2 Phase 3 Corpus collected by the Linguistic Data Consortium. While in the original database two different handsets were used (carbon and electret), in the subset selected for this paper, we only used data from electret handsets.

The database was separated into three subsets: a training set for the world model, and both a development set and an evaluation set of clients. Furthermore, for each client, there was training material and test accesses.

We separated the data into male and female data, in order to create two different world models. The male world model was trained on 137 speakers for a total of 1.5 hours of speech, while the female world model was trained on 218 speakers for a total of 3 hours of speech. After that, the two world models were merged: the new world model have the same mean and variance vectors as the concatenation of the two gender dependent world models and the weights are normalized in order to satisfy the constraint that they should sum to 1.

For both development and evaluation clients, there was about 2 minutes of telephone speech used to train the models and each test access was less than 1 minute long. The development population consisted of 45 males and 45 females, with 417 males and 506 females in the evaluation set. The total number of accesses in the development population was 5135 and 63573 for the evaluation population with a proportion of 10% of true accesses.

4.2. General Methodology

All the experiments described here have followed the same methodology. First, the original waveforms were sampled every 10ms and then parameterized into 16 LFCC coefficients and their first derivative, as well as the energy together with its first derivative, for a total of 34 features.

Afterward, a *bi-Gaussian method* (see section 2.2) was used in the baseline system in order to remove the silence frames from the data. With the new approach, we kept all frames and let the fusion model discard the (non-informative) silence frames.

While the energy was important in order to remove the silence frames, it was not appropriate for the task of discrimination between clients and impostors, and was thus removed from the features after silence removal. Hence, the world and client models were trained with 33 features (instead of 34).

In order to select the various hyper-parameters (such as the number of Gaussians, the MAP factor, etc), we used the following methodology: using the development set, for each value of the hyper-parameter to tune, we trained the client models using the training data available for each client. We then selected the value of the hyper-parameter that optimized the *Equal Error Rate* (EER) on the test accesses of the development set. Finally, we trained the models of the evaluation set using these hyper-parameters and report the results obtained on the test accesses of the evaluation set. Hence, these results are unbiased as the corresponding data has not been used for any purpose during the development of the models.

4.3. Specific Methodology for the Proposed Model

In this case, two discriminant models are created, one for each gender (as we do for world models). First, a 5-fold cross-validation method is used in order to tune the hyper-parameters such as learning rate, number of hidden units, "a priori" decision threshold, etc. Then, given the best hyper-parameters, the model is retrained on the whole training data.

4.4. Experimental Results

The values of the hyper-parameters found on the development set were the following: 128 Gaussians for the world model, the MAP adaptation factor is equal to 0.5, the variance flooring is equal to 60% of the global variance, $\alpha = 0.6$ for the synchronous alignment factor.

Concerning the SVM, we give results for two kernels, a linear kernel and a Gaussian kernel with $\sigma = 800$. The MLP is in fact a simple regression model (no hidden units).

Results of the experiments are given in Table 1. FAR represents the *false acceptance rate* (number of false acceptances divided by number of impostor accesses), FRR is the *false rejection rate* (number of false rejections divided by number of client accesses), while HTER is the *half total error rate* (the average of FAR and FRR). The DET curves of the two best methods are also shown in Figure 1.

Method	FAR	FRR	HTER
Baseline	13	9.72	11.36
Linear SVM	14.57	8.93	11.75
Gaussian SVM	15.3	8.22	11.76
Linear MLP	8.83	12.02	10.42

Table 1. Comparison between baseline (GMMs with MAP), simple linear regression (using an MLP), and SVMs on the evaluation set of the NIST 2001 database when the decision threshold has been selected according to the EER criterion on the development set.

Figure 1 shows that the linear MLP gives overall better results than the baseline system[1]: the DET curve of the MLP is always under the DET curve of the baseline system.

Moreover, the a priori results, given in Table 1, show that the new approach gives statistically significantly better results than the baseline system with a confidence level of 99%[2]. Both the DET curve and the a priori results show that it is indeed possible to replace a silence/speech detector by a discriminant model.

[1]SVMs results are not given, but they are not as good as the Linear MLP and similar to the baseline system.

[2]with a standard proportion test, assuming a binomial distribution for the errors, and using a normal approximation.

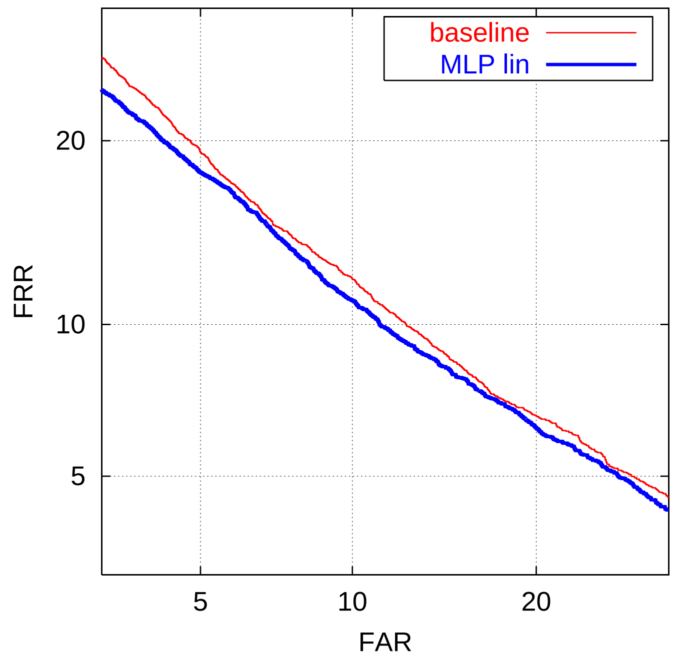

Fig. 1. DET curves on the NIST 2001 evaluation set for the baseline (GMMs with MAP) and simple linear regression (using an MLP).

While getting good performance is important, there are still other reasons to opt for the new model:

- no a priori knowledge is used regarding the discrimination of the input frames, hence the resulting system should be more robust to variable environmental conditions,
- since no frames are removed from the inputs, there is no discontinuity in the data, which can be useful for models that use temporal information (for instance, discriminant models such as MLPs could more easily integrate temporal windows of several consecutive input frames, which is much more difficult to do when there are discontinuities in the temporal sequence),
- this approach can help to better understand GMM models for speaker verification, especially in the way these local models cluster the frames.

5. CONCLUSION

In this paper, we have proposed a new method to replace the usual silence/speech detector in text independent speaker verification systems. A new scoring method was proposed, in order to have one score per pair of Gaussians (world/client). Discriminant models such as MLPs or SVMs were then trained to fuse these scores. The experimental results showed that the new system performed statistically significantly better than the baseline system without having to decide on any prior segmentation of the input data.

Regarding future work, the hypothesis that each Gaussian represents a class of sub-unit of speech should be verified more extensively experimentally, and based on these results, new discriminant models could emerge.

6. ACKNOWLEDGMENTS

The authors would like to thank Ivan Magrin-Chagnolleau for fruitful discussions. This research has been partially carried out in the framework of the Swiss NCCR project (IM)2. All experiments were done using the *Torch* package [6].

7. REFERENCES

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1–3, 2000.

[2] Ivan Magrin-Chagnolleau, Guillaume Gravier, and Raphaël Blouet, “Overview of the 2000-2001 ELISA consortium research activities,” in *2001 A Speaker Odyssey*, June 2001, pp. 67–72.

[3] Johnny Mariéthoz, Dominique Genoud, Frédéric Bimbot, and Chafik Mokbel, “Client / world model synchronous alignement for speaker verification,” in *6th European Conference on Speech Communication and Technology — Eurospeech’99*, Budapest, Hungary, September 5–10 1999.

[4] V. N. Vapnik, *The nature of statistical learning theory*, Springer, second edition, 1995.

[5] C. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[6] R. Collobert, S. Bengio, and J. Mariéthoz, “Torch: a modular machine learning software library,” Technical Report IDIAP-RR 02-46, IDIAP, 2002.