A Multiclass framework for Speaker Verification within an Acoustic Event Sequence system

Nicolas Scheffer, Jean-François Bonastre

LIA, Université d'Avignon
84911 Avignon CEDEX 9, France
{nicolas.scheffer,jean-francois.bonastre}@univ-avignon.fr

Abstract

Building acoustic events and their sequence analysis (AES system) is a method that proved its efficiency in [1]. Indeed, the methodology combines the power of the world model GMM, used in state-of-the-art speaker detection systems, for extracting speaker independent events with an analysis of these event sequences via tools usually used in so-called High Level Speaker Detection systems. The efficiency of this system has been validated at the last NIST evaluation campaign. This paper aims at proposing a new framework by applying an AES system on multiple classes, C-AES. The originality of this work is to consider that intraclass sequence analysis can bring more information than a global analysis on the whole speaker utterance. This paper also proposes a method to take into account the apriori knowledge of the classes within the scoring process. The results support the fact that intraclass information is discriminant for speaker verification, as a combination with a state-of-the-art GMM brings a 12% relative gain at the DCF.

Index Terms: speaker verification, sequential analysis, high-level features, multiclass.

1. Introduction

In the past few years, the automatic speaker recognition field made an extensive use of Gaussian Mixture Models to deal with the problem of text-independent speaker recognition [2]. The latter has fulfilled its role by reducing error rates especially in NIST SRE campaigns[1]. Since the introduction of an extended data task, drastically increasing the amount of data, the community tends to model other types of information, prosodic, phonetics [3], idiolectical [4], usually referred to as high level features [5]. In this area, phonotactic methods [3] have proved to be a promising strategy. It basically assumes that phoneme sequences carry speaker specific information. Its main disadvantage is that it needs a consequent amount of data to be efficient. It also implies that phoneme sequences are the proper time unit to carry out a sequence analysis.
In [1], we proposed a system, named Acoustic Event System (AES), which does not use any apriori information on the type of acoustic events. AES uses the ability of an general/world GMM, the classical UBM, to model any kind of distribution for extracting automatically a set of acoustic events. Using these acoustic events gathered from the UBM model, we perform a sequence analysis, thanks to an Ngram approach. The AES system showed a good level of performance during NIST-SRE 2005 evaluation, when combined with a standard GMM system.
This paper proposes a methodology to carry out a multiclass analysis using an AES system. In the literature, multiclass analysis are widely used in speech recognition, whereas for speaker verification, approaches like [3] and [6] showed relatively good performance.
The C-AES consists in applying an AES system on each classes before combining all information to perform the multiclass analysis. It aims at underlying the importance of the intraclass information compared to the interclass one. The classes are made in the same manner as acoustic events and are called Class Events, as they are at a much lower resolution than the standard events in an AES system. The basic idea is that it brings other information than classical systems which usually work on the whole utterance. Indeed, GMM based systems perform best with a single transformation for each speaker utterance (e.g.:MAP adaptation). We support the hypothesis that characterizing a speaker by several intraclass analysis brings other information. In order to succeed in this approach, the apriori knowledge on the class is important and a modification of the TFLLR (Term Frequency Log Likelihood Ratio) kernel, presented in [7] is needed. This paper is organized as follows. First, a brief description of an AES system is presented in 2 before presenting the principle of C-AES in section 3. In this section, the approach for generating the Class Events can be found and the method to combine multiclass information is presented. Before concluding on this work in 5, section 4 summarizes the different experiments and results, in which a C-AES system and a standard AES system are compared.

2. The LIA Acoustic Event Sequence system

The AES system (presented and validated at NIST 2005) has its principle explained in this section. It consists of two main parts: a speaker independent process where acoustic events are built, and a speaker dependant process during which the verification task is performed.

2.1. Speaker independent acoustic dictionary

The AES dictionary symbols are generated by the GMM system UBM presented in 4.2. The amount of training data is voluntarily high so that information contained in the GMM is maximal. The first step in building the dictionary consists in extracting the Gaussian with maximum likelihood and to use its associated index as a symbol. At first, the dimension of the dictionary generated is equal to the number of GMM components (2048 in this paper).
The next step consists in reducing the number of symbols to form Acoustic Events. The reader can refer to [1] for a more detailed ex-

[1] www.nist.gov/speech

planation of the reduction process. The original dictionary consisting of 2048 words (indexes of the Gaussian) is reduced by clustering Gaussian indexes. A 128 dictionary size, number determined empirically, showed a good level of performance. For further notice, symbols from the AES dictionary are called Feature Events.

2.2. Symbol generation procedure

The adopted strategy is to transform all the parameterized signals into symbol sequences. Indeed, each feature file belonging to the train, test and world model set is submitted to the same process. For a given utterance, each frame is passed through the background model to compute its 1-best component. Then, the index of this component is replaced with its corresponding symbol in the dictionary previously built. It is worth noting that the resulting sequence length is the same as the number of frames in the feature file.

2.3. Speaker specific information modeling

Having produced a speaker independent dictionary, a Bag of Ngram approach is used to build speaker specific models upon this dictionary. Indeed all Ngrams (seen at least twice) of order 1 to 3 are computed for every speaker utterances. We used all Ngrams seen in the background data that have occurred more than 10 times. All probabilities for each Ngram are computed. The modelling and scoring process is performed within a SVM classifier and the TFLLR (Term Frequency Log Likelihood Ratio) kernel framework.

3. C-AES: A multiclass analysis for AES

This section aims at explaining the methodology proposed for a multiclass analysis within the use of an AES framework.

3.1. Principle

The main idea of a multiclass AES system (C-AES) is to perform the sequence analysis inside classes of the speech signal. Such classes can be of two different types:

- Phonetic based classes, e.g.: vowels, fricatives, nasals, . . . ,
- Acoustic based classes obtained in an unsupervised way.

The second approach is chosen in this work, as the classes are here considered as another type of Acoustic Events, but at a much lower resolution. Their generation naturally follows the same procedure seen in (2.1). Building a C-AES system consists then in the following steps, illustrated by figure 1:

- Generation of the Feature Event (FE) set as in a classical AES system. (see 2.1),
- Generation of the Class Event (CE) set (see 3.2),
- Application of an AES system for each CE independently,
- Combination of the information coming from the multiple systems (see 3.4).

3.2. Generation of Class Events

Two different sets of Acoustic Events are generated as explained in 2.1, the Class Events and the Feature Events. The former being at a much lower resolution than the latter. The CE are produced by performing an additive reduction step aiming at clustering the Feature Events into bigger classes. The CE have a different role compared to the FE as no sequential analysis will be performed at the CE's resolution. In this work, the dictionary size for FE is fixed to 128 and to 8 for CE.

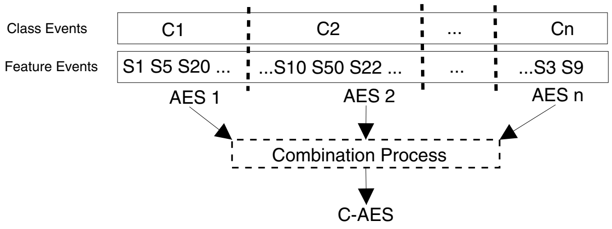

Figure 1: *Combination of multiclass information: an AES system is applied inside each of the n Class Events along their Feature Events. A combination process is performed before the concatenation of all information.*

The process of generating symbols from speech utterances follows the principle explained in 2.2, except that for each utterance two sets of symbol streams are now computed. It is worth noticing that a CE is a group of FEs, which leads to the following remarks:

- This framework can be seen as a sequence selection scheme in a standard AES system, i.e. all interclass sequences are discarded in this framework.
- A C-AES system implicitly maximizes the coverage during the analysis. Indeed, each Class Event has a smaller dictionary size than in a standard AES, the coverage of all possible sequences inside classes is therefore maximize. To illustrate this last point, Table 1 shows the repartition of the dictionary size among the different Class Events.

Table 1: *Feature Event Dictionary size for the 8 Class Event C-AES.*

C_k	1	2	3	4	5	6	7	8
Dictionary size	37	10	11	29	9	9	13	10

3.3. Kernel Construction

The AES system uses a Support Vector Machine framework for the scoring process with the use of the TFLLR kernel. This kernel is used to produce feature vectors for Ngram type approaches. For our purpose, the TFLLR kernel does not entirely fulfil our needs. Well-known normalization techniques such as Rank Normalization or Z-Norm could help to normalize out class influence but do not fit our problem. Indeed, they are powerful when no apriori knowledge is available. In section 4, results tend to support the fact that this information is crucial, either is the way of estimating it.

The kernel needs to be modified slightly so that it can perform a multiclass scoring, by taking into account the CE influence. Let us consider tokens k belonging to a bag-of-Ngram. Let the token k likelihood on a data sequence X be defined as $p(k|X)$, the TFLLR kernel is defined by:

$$\sum_k \frac{p(k|X_1)}{\sqrt{p(k|X_W)}} \frac{p(k|X_2)}{\sqrt{p(k|X_W)}} - 1 \tag{1}$$

where X_1,X_2,X_W are the respective training data of two different speakers and the background model. The kernel construction

finally resides in the weighting of speaker likelihoods by the likelihood of the background model.
All tokens k are Ngrams, whose symbols belong to the Feature Event dictionnary. Let C_k being a Class event related to the token k, the computation of the LLR for the whole utterance is given by:

$$\ell\ell r(k|X) = \sum_k p(C_k|X)\ell\ell r(k|X, C_k) \tag{2}$$

In order to take into account the information imbalance between classes, the following mapping (i.e. producing a fixed-dimension vector from a token k and a Bag of Ngram B, as an input of a SVM classifier) is used:

$$\phi(k, X) = p(C_k|X)\frac{p(k|X, C_k)}{\sqrt{p(k|X_W)}}, \forall k \in B \tag{3}$$

The resulting feature value is the token probability weighted by the background model probability and by its Class event probability

3.4. Multiclass information combination and SVM modelling

The process of construction of an input vector for a SVM classifier from an utterance is presented. For each Class event, a sequence analysis is made inside the class with its FE related dictionary. This produces 8 vectors in our case. For a single utterance, all vectors from all Class Events are then concatenated after being weighted by the class probability described in the previous paragraph.

In order to build impostor models, all speakers used to train the world model have been used to represent the negative labelled data. The input of the classifier is the concatenation of all impostor trials and the target speaker trial. The maximum margin decision is found by passing this input through a linear kernel. We used the SVM-Light toolkit by Thorsten Joachims [8] to induce SVMs and classify instances. To compensate for the severe imbalance between the target and background data, we adopt a cost model to weight the positive examples 200-fold with respect to the negative examples (a number found empirically). The scores obtained in this manner are then normalized using Tnorm.

4. Experiments and Results

In this section, we first present the protocol used for the experiment based on the NIST SRE evaluations. Next, the baseline GMM/UBM system used for experiment is described. Results of the C-AES system are presented as well as different methods to estimate the class weighting factor precised in eq. 3. To conclude, the performance when combined with a standard GMM-UBM system is compared with the classical AES system.

4.1. Datasets

Speaker verification experiments, presented in section 4 are performed based upon the NIST 2005 database, all trial set (det1), male speakers only. This condition consists of 280 speakers. Train and test utterances contain 2.5 minutes of speech in average (telephone conversation). The whole speaker detection experiment consists in 13624 tests (951 target tests). Each test is made independently and the use of information from other tests to take a decision on the current test is forbidden. Results are given as detection cost function (DCF) and equal error rate (EER). DCF is the Bayesian risk function defined by NIST with $P_{target} = .1$, $C_{fa} = 1$, and $C_{miss} = 10$, as well as Detection error tradeoff (DET) curves [9].

4.2. The LIA_SpkDET UBM/GMM system

The background model used for the experiments is the same as the background model used by the LIA for the NIST SRE 2005 campaign (male only). The training is performed based on NIST SRE 1999 and 2002 databases, and consists in 1.3 million of speech frames. Training was performed using the ALIZE and LIA_SpkDet toolkits[2] [10]. Frames are composed of 16 LFCC parameters and its derivatives. A normalization process is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance. The background model posses 2048 components and no component variance is above 0.5. The speaker model parameters are obtained by adapting the world model mean parameters [2]. The reader will find in section 4 its performance with a TNormalisation applied on scores.

4.3. Estimation of class probabilities

This section investigates two approaches for the class probability estimate. The first one consists in taking the apriori probability for each Class Event, the other makes the use of MAP estimation for this probability.

4.3.1. Aprori class as a weighting factor

To estimate the apriori probability of each Class Event, all Class Event data from the background model has been used, i.e. $P(C_k) = P(C_k|X_W)$. The probability is the CE's frequency of occurrence in the data. In our case, Table 2 gives the weighting factor used for the experiment.

4.3.2. Using MAP for the estimation

A maximum a posteriori (MAP) approach can also be employed to estimate this probability. The estimation on a single utterance not being enough precise, MAP enables to rely on apriori probability if the class is not present in the utterance. Precisely, if $\hat{p}(C_k|X)$ is the new probability estimate, then:

$$\hat{p}(C_k|X) = \alpha p(C_k|X) + (1-\alpha)p(C_k|X_W), \text{ with } \alpha = \frac{C(k)}{C(k)+\tau}$$

where $C(.)$ is the count operator. τ is called regulation factor (usually found empirically). Here, it has been found by dividing the average utterance length by the number of class (with 8 classes τ has been fixed to 1000). Table 3 shows the effect of this technique on the performance of the C-AES system.
The results tend to prove the fact that the integration of information from each class is mandatory. Indeed, the system without any weighting performs two times worse than the one with the apriori estimation. The second experiment goes one step further by showing that the estimation method is also very important. An absolute gain of 1% both at the EER and the DCF is observed when MAP estimation is used.

4.4. System combination

From the results obtained at the last paragraph, it is clear that a C-AES system cannot compete with a standard GMM/UBM. How-

[2] http://www.lia.univ-avignon.fr/heberges/ALIZE/

Table 2: *A priori class probability for the 8-class model.*

C_k	1	2	3	4	5	6	7	8
$P(C_k)$	.41	.02	.11	.26	.07	.03	.06	.03

Table 3: *Performance of C-AES with different estimates of CE probability*

System	DCF (x100)	EER
No weighting factor	9.80	28.5
A Priori	6.46	14.87
MAP estimation	5.87	13.89

ever, it seems particularly efficient in the fusion process as we present some experiment on system combination below.

For these results, all systems have been Tnormed and an arithmetic mean between systems is performed. A GMM/UBM is presented and corresponds to the baseline presented in 4.2. Table 4 presents the result of different combinations between baselines and the C-AES system, while Figure 2 illustrates the results in terms of a DET curve (fusion weights have been found empirically to optimize the DCF).

While a C-AES system performance is slightly lower than the classical AES system, it is interesting to notice that a better performace is observed in fusion. Indeed, a relative gain of 12% and 7% is observed at the DCF and EER respectively compared to 7% and 6% for a standard GMM/UBM and AES fusion. This methodology tends to support the proposition that the important information for a sequential analysis resides inside the classes of the speech signal and that interclass relationships is less important for such a dynamic related analysis.

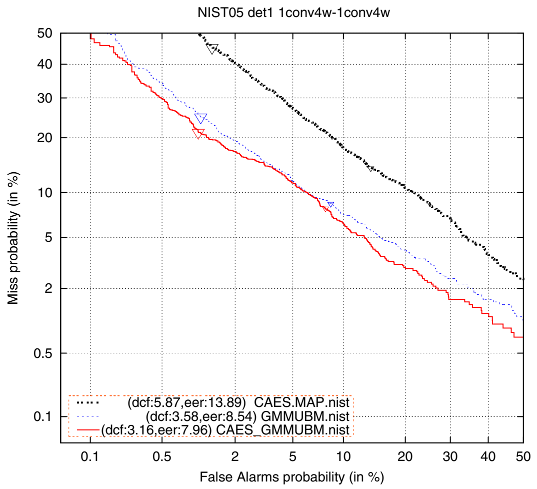

Figure 2: *DET curves for baseline, CAES, and combined systems.*

5. Conclusion and Future Work

By proposing a framework in section 3 for applying sequential analysis in a multiclass framework, this paper supports the idea that multiple analysis for a single utterance contains speaker-specific information. By extending the TFLLR kernel so that scoring on different classes can be performed, we proposed a framework for combining the different class information. We also showed in section 4 that estimation of prior knowledge has to be carefully made.

Future work will focus on a extending this framework to a multi resolution framework by generating a lot more segmentation sets and feature sets in different resolutions in order to combine these multiple analysis into a single vector for the classifier.

Table 4: *Combination of different systems*

System	DCF (*100)	EER
GMM	3.58	8.54
AES	5.37	13.33
CAES	5.87	13.89
GMM + AES	3.31	8.04
GMM + CAES	3.16	7.96

6. References

[1] N. Scheffer and J-F. Bonastre, "Speaker detection using acoustic event sequences," in *INTERSPEECH Conference, Lisboa, Portugal*, 2005.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] W. D. Andrews, M. A. Kohler, J. P. Campbell, and J. J. Godfrey, "Phonetic, idiolectal, and acoustic speaker recognition.," in *Odyssey Conference, Chania, Greece*, 2001.

[4] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *EUROSPEECH Conference, Aalborg, Denmark*, 2001, pp. 2521–2524.

[5] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition.," in *EUROSPEECH Conference, Geneva, Switzerland*, 2003, pp. 2665–2668.

[6] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and Venkataraman A., "Mllr transforms as features in speaker recognition," in *INTERSPEECH Conference, Lisboa, Portugal*, 2005.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R Leek, "High-level speaker verification with support vector machines," in *ICASSP Conference, Montreal, CANADA*, May 2004, pp. 73–76.

[8] T. Joachims, "Making large-scale svm learning," in *Practical. Advances in Kernel Methods - Support Vector Learning,B. Schokopf and C. Burges and A. Smola, MIT Press*, 1999.

[9] A. F. Martin and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 97)*, Septembre 1997, pp. 1895–1898.

[10] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005)*, Philadelphia, USA, March 2005.