ABSTRACT
The paper describes a structural framework for the design of a speaker recognition system based on multiple models. This combination is not only at the recognition level, but also at a joint training of the models. This unified training of the models uses a common structure: a decomposition tree of the set of data of normalization speakers. For the experiments, the Gaussian Mixture Model and the Auto-Regressive Vectorial Model are the two models we have selected to test the structural framework of the speaker verification scoring combination. This approach has been tested on a subset of the 30"-NIST'97 Speaker Recognition Evaluation corpus. The list of the files of this subset (i.e., normalization, training and test) can be found at http://www-apa.lip6.fr/PAROLE/ICSLP2000/.

1. INTRODUCTION
Many models have been investigated for the development of speaker recognition systems [1] (e.g., Vector Quantization (VQ), the Gaussian Mixture Models (GMM), Sub-Band GMM, Hidden Markov Model (HMM), Predictive Models such as Auto-Regressive Vectorial Models (ARVM), Neural Networks). For the improvement of the performance of the speaker recognizers, one promising investigation we believe is to combine some of these methods expecting a gain when they are not correlated. This combination may be processed at a symbolic level (i.e., combination of the symbolic recognition) or at a numeric level (combination of recognition scores). Besides the usual problems of normalization in relation with the imposters modeling, the heterogeneous nature of the chosen models pose other problems of normalization and possible bias making difficult as well the comparisons as the combination of the scores obtained from the various methods.

In previous work [2] where we have investigated the cooperation and the competition of models, the combination of analytic recognition (i.e., final score computed from the score recognition of each acoustic vector of the utterance) was shown to be more efficient than the combination of segmental recognition (i.e., final score computed for a whole sequence of acoustic vectors of the utterance). Among the various models, some of them are unfit to an analytic recognition (i.e., the elementary result of the recognition process isn’t related to a single vector). Indeed, some models are typically designed for segmental recognition which makes difficult their use at an analytic level (e.g., HMM), or are too computation time consuming (e.g., ARVM) for the purpose of an analytic recognition. The fusion of scoring is a challenging problem: i) the complexity of a principle of decision, related to the use of a model, increases with the variability of the data, ii) the efficiency in combining the individual method scores decreases while the combination complexity of the scores increases.

As a function of the chosen models, a problem can arise for the numeric combination of scores: the nature of the similarity measures computed between a sequence of acoustic observation vectors and a given model. This normalization problem is posed, for instance, when combining models with the similarity measures computed as i) likelihood (e.g., GMM), and ii) inter-models measures (e.g., ARVM). A simple linear combination of scores isn’t always efficient to take advantage of the complementarity of the methods.

To overcome these problems, we propose a clustering of the representation space in terms of neighborhood of the observations and to define a principle of decision for each cluster. The clusters are the common structure used for the training of each of the models. For the following experiments, the representation space has been clustered according to the same principle like used in GMM training, the GMM being one of the two methods we have chosen to combine. From this common structure of data for the training of the models, we expect for a bias as low as possible at comparing and at combining the verification scores of the models.

In section 2 and section 3 we present the two models we have chosen to combine: the ARVM and the GMM. Section 4, describes the basis of the experimental protocols, i) the NIST SRE based corpus we have delimited, ii) the normalization of the different similarity measures we have adapted for imposters and target speakers modeling, and iii) the resulting Equal Error Rate (EER) obtained independently for each of the models. In section 5, the structural framework we propose for combining the methods is defined: the joint training of the models based on the
On the contrary of the GMM, an universal ARVM isn’t a sequence of observation avoiding the computation of models distance. The similarity measure is then computed as an inter-models similarity measure between a sequence of observations and a model of the sequence is first computed. The similarity measure between a sequence of observations and a speaker ARVM, the model of the sequence is first computed. To compute the similarity measures as used with ARVM. If the measure is informative between two “close” speakers, the measure is less efficient. For each target speaker, it is necessary to train speaker-dependent ARVM.

2. GMM - ANALYTICAL MODELS

A GMM $\Lambda_{GMM}$ is basically an analytic model. The distribution of a $p$-dimensional acoustic vector $o$ (i.e., an observation) is computed as a linear combination of Gaussian densities.

$$p(o/\Lambda_{GMM}) = \sum_{i=1}^{m} w_i G_i(o) \quad \text{with} \quad \sum_{i=1}^{m} w_i = 1 \quad (1)$$

with mixture weights $\{w_i\}_{i=1}^{m}$ and Gaussian densities $\{G_i(o)\}_{i=1}^{m}$:

$$G_i(o) = \frac{1}{\sqrt{(2\pi)^p \text{Det}(\Sigma_i)}} e^{-\frac{1}{2} (o-\mu_i)^T \Sigma_i^{-1} (o-\mu_i)} \quad (2)$$

with $(\mu_i, \Sigma_i)$ the mean vector and the covariance matrix of the Gaussian density. The similarity measure between a sequence of observations $O = \{o_t\}_{t=1}^{T}$ and a GMM model $\Lambda_{GMM}$ is computed as the average likelihood of the sequence:

$$L(O/\Lambda_{GMM}) = \frac{1}{T} \sum_{t=1}^{T} p(o_t/\Lambda_{GMM}) \quad (3)$$

A speaker independent GMM is trained on a large number of speakers (i.e., set of normalization speakers) aiming at creating an Universal Background Model $\Lambda_{UBM}$ of the imposters (UBM). The target speakers GMM are adapted from the UBM.

3. ARVM - SEGMENTAL MODELS

A speaker ARVM of order $q$ is basically a segmental model. At time $t$, an observation $o_t$ (i.e., a $p$-dimensional vector) is predicted from its $q$ temporal predecessors:

$$o_t = \sum_{i=1}^{q} A_i o_{t-i} + e_t \quad (4)$$

with $\{o_t\}_{t=1}^{T}$ a sequence of observations, $\{A_i\}_{i=1}^{q}$ the set of $q$ prediction matrices $(p, p)$; $\{e_t\}_{t=1}^{T}$ a vectorial white noise of covariance matrix $E$. The coefficients of matrices $A_i$ are estimated in order to minimize the trace of the matrix $E$. To compute the similarity measure between a sequence of observations and a speaker ARVM, the model of the sequence is first computed. The similarity measure is then computed as an inter-models measure based on a likelihood ratio between the errors of prediction of the models. A fast computation of this inter-models distance $[2]$ uses the $q+1$ correlation matrices of the sequence of observation avoiding the computation of $\{e_t\}_{t=1}^{T}$. On the contrary of the GMM, an universal ARVM isn’t efficient. For each target speaker, it is necessary to train speaker-dependent ARVM.

4. EXPERIMENTAL PROTOCOLS

4.1. NIST SRE based experimental corpus

Experiments are conducted on a subset of the 1997 NIST Speaker Recognition Evaluation (SRE) corpus [4]. The NIST SRE corpus is derived from Switchboard-2 phase1 conversational telephone corpus. The recordings of the target speakers are calibrated into one minute duration. The recordings are composed of several short speech segments (mean of 1.5 seconds and a standard deviation of 1.2 second) uttered in a conversation on telephone. For each target speaker, four recordings of speech data are distinguished: i) one “reference” recording is taken from a conversation on telephone (i.e., a given conversation and a given handset), ii) one recording is taken from the same “reference” conversation (i.e., same conversation and same handset), iii) one recording of a different conversation and same handset, iv) one recording of a different conversation and different handset. A target speaker is trained from spoken data of two minutes duration (i.e., a set of two recordings). Taking two of the four types of recording, three training conditions are defined as: i) one-session (1&2), ii) one-handset (1&3), ii) two-handset (1&4).

We have chosen the one-handset training condition with electret type handset. For the experiments, we have i) a set of normalization (for each gender, the first fifty target speakers), ii) a target speaker set (for each gender, the 50 following target speakers). The 1997 NIST SRE corpus test consists in target and test speakers from the same gender. From the original test set (27,003 male tests, 27,500 female tests), we have chosen as test set $S$: for each gender, the first 2,500 tests whose target and test speakers have used electret handset and whose speakers are not belonging yet to the set of normalization. From this test set, two subsets are distinguished according to the phone numbers of the test speaker and the target speaker: i) same number ($S1$), ii) different numbers ($S2$).

4.2. Normalization

Investigated in statistical pattern recognition, a challenging problem is posed for speaker verification: the determination of appropriate decision threshold. Normalization of the scores [5] is a necessary step for speaker verification systems. Among the investigations on the scores normalization, two main principles are distinguished. The first one consists in subtracting a statistic of the similarity measures between the test and the models of normalization speakers from the similarity measure between the test and the target model. As examples of statistics yet proposed: the highest one, the mean, the first $n$ highest. The second principle consists in building a model of the set of normalization speakers (e.g., UBM) and in subtracting the similarity measure between the test and this model from the similarity measure between the test and the target model.

These methods of normalization aren’t well-fitted in case of similarity measures as used with ARVM. If the measure is informative between two “close” speakers, the measure is less...
informative for very “far” speakers. As shown on Figures 1&2, it’s difficult to decide what kind of statistic is more adapted.

**Figure 1:** Intra-speaker measures histogram

**Figure 2:** Inter-speakers measures histogram

To normalize an inter-models measure $D$, two sets of normalization measures are used. Related to a set of $N$ speakers of normalization, these sets aim both at representing the speakers variability (i.e., intra-speaker and inter-speakers). The first one $\{D_s^T\}_{s=1,...,N}$ is a set of intra-speaker measures computed, only once, for each speaker of normalization. The second one $\{D_s^I\}_{s=1,...,N}$ is a set of inter-speakers measures computed, for each test utterance, as the measure between the test and a speaker of normalization. Let these two sets of normalization measures be increasing sorted. Two normalized measures will be computed from these two sets of measures.

The first normalized measure $P^T(D)$ is interpreted as the estimation of the probability the measure is an intra-speaker measure, and consequently the test speaker is the target speaker.

The second normalized measure $P^I(D)$ is interpreted as the estimation of the probability the measure is an inter-speakers measure, and therefore the test speaker is an imposter.

The two normalized measures are deducted from the rank $r$ and the relative position $r_p$ of the measure $D$ within the range of the sets of normalization measures by the following formula:

$$P^T(D) = \begin{cases} 1 & \text{if } D < D_s^T \\ 0 & \text{if } D > D_s^N \\ \left(\frac{N-r-r_p+1}{N}\right) & \text{if } D_s^T < D < D_{s+1}^T \end{cases}$$  

$$P^I(D) = \begin{cases} 1 & \text{if } D < D_s^I \\ 0 & \text{if } D > D_s^N \\ \left(\frac{r_r+r_p}{N}\right) & \text{if } D_s^I < D < D_{s+1}^I \end{cases}$$

For example, a similarity measure equal to 0.3 will give two normalized scores according to the histograms of Figure 1&2 : $P^T(0.3)=0.02$ and $P^I(0.3)=0.67$.

### 4.3. ARVM and GMM Results

For all the experiments, the same parameters (12 Cepstral coefficients) are used for the GMM and the ARVM. The order $q$ of the ARVM equals 2. The number $m$ of Gaussian density functions equals 64.

The task of speaker verification is to determine whether a specified target speaker $T$ is speaking during a given speech sequence $O$.

The speaker verification scores for the GMM $P_{GMM}$ and for the ARVM $P_{ARVM}$ are computed as follows:

$$P_{GMM} = L(O/\Lambda_{GMM}^T)/L(O/\Lambda_{GMM})$$

where the target GMM $\Lambda_{GMM}^T$ is adapted from the Gaussian UBM $\Lambda_{GMM}^{UMB}$ using a Maximum A Posteriori (MAP) algorithm.

$$P_{ARVM} = \alpha P^T(D, \Lambda_{ARVM}^T) + \beta P^I(D, \Lambda_{ARVM})$$

with a weighting factor $\alpha$ allowing the combination of the two normalized measures of $D$. The value $\alpha$ must be trained on a development corpus in order to minimize a quality criterion (e.g., Equal Error Rate).

Table 1 gives the results for a weighting factor $\alpha$ equal to 0.

<table>
<thead>
<tr>
<th>Basic Models</th>
<th>ARVM</th>
<th>GMM</th>
<th>ARVM</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>10.8%</td>
<td>12.1%</td>
<td>16.8%</td>
<td>23.4%</td>
</tr>
<tr>
<td>S1</td>
<td>3.5%</td>
<td>5.8%</td>
<td>5.5%</td>
<td>7%</td>
</tr>
<tr>
<td>S2</td>
<td>15.8%</td>
<td>17.4%</td>
<td>24%</td>
<td>34.3%</td>
</tr>
</tbody>
</table>

**Table 1:** Equal Error Rate (EER) over the three distinguished test sets for the basic models : GMM and ARVM.

We can remark : i) the EER for the GMM is higher than the EER for the ARVM, ii) for both of the models, the EER for females are much higher than for males. An explanation is the low number of Gaussian densities ($m=64$) and a high variability of female speech over these recordings (i.e., emotional state modifications and many laughter).

Table 2 gives the EER for the optimal value of $\alpha$.

<table>
<thead>
<tr>
<th>Optimal $\alpha$</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>10.6%</td>
<td>16.6%</td>
</tr>
<tr>
<td>S</td>
<td>10.6%</td>
<td>16.6%</td>
</tr>
</tbody>
</table>

**Table 2:** Optimal Equal Error Rate (EER) for the basic model ARVM with the optimal weighting factor $\alpha$.

We remark that the combination of the probabilities of target $P_{ARVM}(D)$ and of non-imposter $(1-P_{ARVM}(D))$ doesn’t improve significantly the EER. An explanation is that the set of intra-
5. GMM AND ARVM COMBINATION

5.1. Combination Principle of Models

The combination of models we have designed is based on the following principle: several speaker models are expected to be easier to combine when they are specialized on a same strategic part of the acoustic space. Many studies [6] have aimed at the determination of the most discriminant acoustic spaces in terms of speaker recognition as well as the determination of the models the most adapted to a precise neighborhood (e.g., phonetic clusters, frequency bandwidth).

In our case, a set of NC acoustic clusters is first defined. The models to combine are afterwards specialized for each of these clusters. NC speaker verification scores are computed per kind of model. For each cluster, a combination of the scores is then computed. At last, either the whole NC scores or a part of these scores are combined to give the final scoring.

There are three steps for the models combination algorithm: the branch step, the bound step and the combination step. At the branch step, each observation of the sequence O is assigned to one of the NC clusters. At the end of this process, the set of observations associated to each cluster \{O_i | i=1,...,NC\} can be represented as a set of speech segments. The choice of the cluster must be guided by the following considerations: i) the mean duration of the segments has to be high enough to allow the training of ARVM of order q; ii) to allow a good specialization of the models, the size of the clusters has to be not too high. For our first experiment, the clusters are defined from the following functions of decision \{C_j | j=1,...,NC\} of a Gaussian UBM:

\[
C_j(o) = \sum_{i=1}^{i=NC} w_{ij} G_{ij}(o)
\]  

(9)

The bound step is processed from the sets \{O_{i,c} | i=1,...,NC\}:

\[
O_{c}(O) = \left\{ q_{ij} | c = \arg \max \left\{ C_j(o_i) \right\} \right\}
\]  

(10)

For the ARVM, the utterances of the target speaker and the utterance of the test are both clustered giving the following ARVM scores:

\[
P^T_{ARVM,c} = P^T(D(O,c), \Lambda^T_{ARVM,c})
\]  

(11)

\[
\overline{P}^T_{ARVM,c} = \overline{P}^T(D(O,c), \Lambda^T_{ARVM,c})
\]  

(12)

For the GMM, only the test segment is clustered. The MAP adaptation of the target model can be considered equivalent to the clustering. GMM scores are the following:

\[
P^T_{GMM} = L(O_t, \Lambda^T_{GMM}) / L(O_c, \Lambda^T_{UBM})
\]  

(13)

The final scoring is computed as follows:

\[
S^{NC}_{ARVM\&GMM} = \beta_0 \left[ \alpha_0 P^T_{ARVM} + P^T_{GMM} \overline{P}^T_{ARVM} \right] + \sum_{c=1}^{NC} \beta_c \left[ \alpha_c P^T_{ARVM,c} + P^T_{GMM} \overline{P}^T_{ARVM,c} \right]
\]  

(14)

The optimal weights /α_c, β_c/c=0,...,NC/ have to be trained on a development corpus.

5.2. ARVM&GMM Combination

Preliminary Results

For this first experiment, the value of NC is chosen equal to 2. Table 3 gives the EER for the optimal weights.

<table>
<thead>
<tr>
<th>Optimal Weights</th>
<th>α_0</th>
<th>β_0</th>
<th>α_1</th>
<th>β_1</th>
<th>α_2</th>
<th>β_2</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (S)</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>0.4</td>
<td>9.3%</td>
<td></td>
</tr>
<tr>
<td>Female (S)</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>16.3%</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Optimal Equal Error Rate (EER) for the ARVM&GMM combination.

Though the EER have to be validated on a new test corpus, the results are interesting. We remark: i) the fusion ARVM&GMM combination gives a significant decrease of the EER, ii) the behavior is very different according to the gender of the speaker. For males, the null value of β_0 confirm our hypothesis of a better efficiency of the models combination when the models are specialized on strategic clusters. On the contrary, for females the null values of de β_1 and β_2 weaken temporarily this hypothesis. Like we have yet noticed it, the high variability of the female recordings could explain an insufficient specialization of the clusters.

6. CONCLUSION

A structural framework designed to combine heterogeneous speaker recognition methods has been proposed. We have shown that the complementarity between ARVM and GMM recognizers can be used to combine the advantages of the two methods. Our future work will be the study of the optimal clusters selection with an assessment on the NIST’98 and TeraSpeech [7] databases.

7. REFERENCES

4. www.itl.nist.gov/iaui/894.01/sp_v1p1.htm (NIST'97 Speaker Recognition Evaluation plan)

