Speaker Discrimination Based on a Fusion Between Neural and Statistical Classifiers

Siham Ouamour(✉) and Halim Sayoud

USTHB University, Algiers, Algeria
{siham.ouamour,halim.sayoud}@uni.de

Abstract. Speaker discrimination consists in checking whether two (or more) speech segments belong to the same speaker or not. In this framework, we propose a new approach developed for the task of speaker discrimination, this approach results from the fusion between a neural network classifier (NN) and a statistical classifier, this fusion is obtained once by combining the scores of the simple classifiers weighted by some confidence coefficients and another time, by using the scores of the statistical classifier as an additional input of the Multi-Layer Perceptron (MLP), in order to optimize the NN training (Hybrid model).

In one hand, we notice that the fusion has improved the results obtained by each approach alone and in the other hand we notice that the fusion using the sum of weighted scores, obtained by each classifier alone, seems to be better than the hybrid method. The experiments, done on a subset of Hub4 Broadcast News database, have shown the efficiency of that fusion in speaker discrimination, where the Equal Error Rate (EER) is about 7 %, with short segments of 4 s only.

Keywords: Speaker discrimination · Fusion · Speech processing

1 Introduction

Speaker discrimination (*by voice*) represents an important field in biometry, since the voice remains the unique method used at distance (via telephone). This particularity, has given to speaker discrimination a great importance, especially in secure applications which require very high accuracy. Speaker discrimination consists in checking whether two different pronunciations (*speech segments*) are uttered by the same speaker or by two different speakers. One means used to compare the utterances is to extract the vocal characteristics from each segment, in order to detect the degree of similarity between them.

Speaker discrimination has applications in several domains, like speaker verification, biometry, multimedia segmentation and speaker based clustering.

Different approaches were developed for this purpose, among those two approaches are investigated in this paper: a neural network and a 2^{nd} order statistical measure, but we also propose two other approaches based on the association between the two previous classifiers.

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 213–221, 2016.
DOI: 10.1007/978-3-319-33618-3_22

These different approaches are evaluated on a sub-set of Broadcast News (1996) [1] and our results show that this fusion is really interesting.

2 Some Techniques Related to Speaker Discrimination and Parameterization

Several techniques were developed for the task of speaker discrimination, like GMM (Gaussian Mixture Models) [2], NN (Neural Networks) [3], statistical measures [4], HMM (Hidden Markov Models) [5] …etc. In our research work, we have approached the discrimination problem with four methods; MLP (Multi-Layer Perceptron), statistical measures, Hybrid method and the fusion based on the sum of weighted scores. These different methods are described below.

For the parameterization, we used 37 MFSC coefficients (Mel Frequency Spectral Coefficients) obtained from the calculation of the energies in the mel spectral scale [6, 7]. This dimension has been chosen after a thorough investigation done on the optimal spectral resolution [8, 9].

2.1 Statistical Method

One of the referential methods used for the task of speaker discrimination is the statistical measure of similarity (μ_{Gc}) which is based on the covariance matrix. The statistical measure is used in order to determine the similarity degree (with regards to speaker's features) between the different speech segments.

We recall bellow the most important properties of the approach [10, 11].

Let $\{x_t\}_{1 \le t \le M}$ be a sequence of M vectors resulting from the P-dimensional acoustic analysis of a speech signal uttered by speaker $\boldsymbol{x}$. These vectors are summarized by the mean vector $\bar{x}$ and the covariance matrix X:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \tag{1}$$

and

$$X = \frac{1}{M}\sum_{t=1}^{M} (x_t - \bar{x})(x_t - \bar{x})^T \tag{2}$$

Similarly, for a speech signal uttered by speaker $\boldsymbol{y}$, a sequence of N vectors can be $\{y_t\}_{1 \le t \le M}$ extracted.

By assuming that all acoustic vectors extracted from the speech signal uttered by speaker $\boldsymbol{x}$ are distributed like a Gaussian function, the likelihood of a single vector y_t uttered by speaker $\boldsymbol{y}$ is:

$$G(y_t/\mathbf{x}) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{(1/2)(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \tag{3}$$

If we assume that all vectors y_t are independent observations, the average log-likelihood of $\{y_t\}_{1 \le t \le M}$ can be written as

$$\overline{L}x(y_1^N) = \frac{1}{N}\log\, G(y_1 \cdots y_N|\boldsymbol{X}) = \frac{1}{N}\sum_{t=1}^{N} \log\, G(y_t|\mathbf{x}) \tag{4}$$

We also define the minus-log-likelihood $\mu(\mathbf{x},y_t)$ which is equivalent to similarity measure between vector y_t (uttered by $\boldsymbol{y}$) and the model of speaker $\boldsymbol{x}$, so that

$$Arg\max_x\ G(y_t/\mathbf{x}) = Arg\min_x\ \mu(\mathbf{x},y_t) \tag{5}$$

We have then:

$$\mu(\mathbf{x},y_t) = -\log\, G(y_t/\mathbf{x}) \tag{6}$$

The similarity measure between test utterance $\{y_t\}_{1 \le t \le M}$ of speaker $\boldsymbol{y}$ and the model of speaker $\boldsymbol{x}$ is then

$$\begin{aligned}\mu(\mathbf{x},\mathbf{y}) = \mu(\mathbf{x},\, y_1^N) &= \frac{1}{N}\sum_{t=1}^{N}\mu(\mathbf{x},y_t) \\ &= -\bar{L}x(y_1^N)\end{aligned} \tag{7}$$

After simplifications, we obtain

$$\begin{aligned}&\mu(\mathbf{x},\mathbf{y}) = \\ &\frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1}) + (\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right] - 1\end{aligned} \tag{8}$$

This measure is equivalent to the standard Gaussian likelihood measure (asymmetric μ_G) defined in [8].

A variant of this measure called μ_{Gc} is deduced from the previous one by assuming that $\bar{y} = \bar{x}$ (inter-speaker variability of the mean is negligible).

Thus, the new formula becomes:

$$\mu_{GC}(\mathbf{x},\mathbf{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1})\right] - 1 \tag{9}$$

2.2 Neural Approach for Speaker Discrimination

Knowing the high discriminative capacities of the NNs (neural networks) [12], we opted for the use of a MLP (Multi-Layer Perceptron) with one or two hidden layers and with only one output. Experiments are done on audio signals, of three or four seconds each and extracted from Hub-4 Broadcast News.

The goal of this neural network [13] is to discriminate the different speakers by their speech signals. For this purpose, an input vector extracted from the MFSC coefficients is used.

The NN must have at its input a number of receptive cells equal to the dimension of the example vector [7]. Thus, in case of using a vector with N MFSC coefficients [6, 7], the number of input receptive cells is equal to $2.N$ (corresponding to two different utterances).

The training is performed by the back-propagation algorithm and the NN output will give then an indication on the correlation between the two utterances:

- *If* $NN_{OUTPUT} = 0$ *then it is the same speaker,*
- *If* $NN_{OUTPUT} = 1$ *then the speakers are different,*

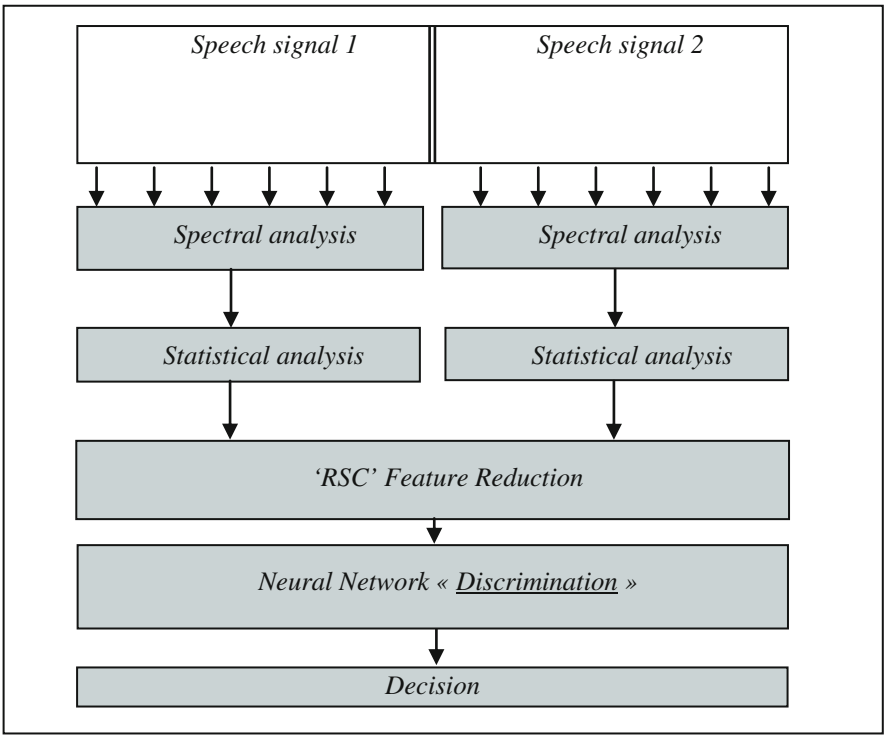

Fig. 1. Comparison between 2 utterances and discrimination decision

If the two segments (utterances) have different characteristics (characterization of the speaker), then we can affirm that these segments belong to the same speaker, otherwise, these segments belong to two different speakers.

Concerning the acoustical-spectral analysis of the signal, a segmentation by windows of 35 ms (ensuring the stationarity) is used in each segment where a spectral analysis is made, in giving one series of MFSC vectors for each segment [6, 7].

This vectors set goes through a statistical process which allows extracting the covariance diagonal elements in each segment. Thereafter, a feature reduction is applied by using a RSC or Relative Speaker Characterization (see section C). These elements are directly injected to the input of the NN which will decide whether the two segments belong to the same speaker or not: see Fig. 1.

2.3 Hybrid Method

Since it has been proved that NNs have an excellent discriminative property, we thought to mix the statistical measure with the neural inputs in order to improve the NN performance: this is the hybrid method.

Thus, a new input is added to the NN, into which we inject the discrimination result given by the statistical measure for each couple of segments, with the corresponding segments and then the training of the NN with this new input is performed as shown in Fig. 2 below.

The hybrid method is summarized as follows: First, the features are extracted from the two segments, then; the statistical measure μ_{Gc} is computed and injected to the NN together with the reduced features, called RSC (*Relative Speaker Characteristic*) [14]. The training is then enhanced by the information brought by the statistical approach.

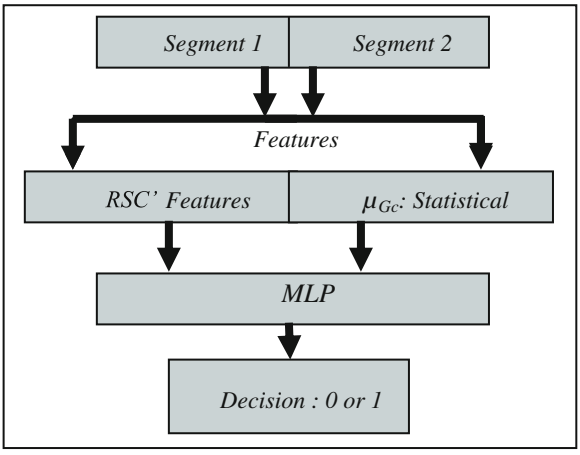

Fig. 2. The hybrid method.

2.4 Fusion

In order to enhance the discrimination performance, we usually use several classifiers which are combined in order to get a better precision: this combination is called Fusion. The fusion in the broad sense can be performed at different hierarchical levels or processing stages. A very commonly encountered taxonomy of data fusion is given by the following three-stage hierarchy [15, 16]:

(a) *Feature level* where the feature sets of different modalities are combined. Fusion at this level provides the highest flexibility but classification problems may arise due to the large dimension of the combined (concatenated) feature vectors.
(b) *Score (matching) level* is the most common level where the fusion takes place. The scores of the classifiers are usually normalized and then they are combined in a consistent manner.
(c) *Decision level* where the outputs of the classifiers establish the decision via techniques such as majority voting. Fusion at the decision level is considered to be rigid for information integration.

In our case, we chose the fusion at the score level.

If the simple scores are denoted by S_j, then the fusion score Sf is given by:

$$Sf = \sum_{j=1}^{N} C_j S_j \tag{10}$$

where C_j represents the weighing coefficient (confidence) for the classifier "j" and N denotes the classifiers number.

With

$$\sum_{j} C_j = 1 \tag{11}$$

and $C_j \in [0.1, 0.9]$

The coefficient C_j represents the relevance of the classifier j.

3 Results and Discussion

The audio database, used in our experiments, is an extract of Broadcast News "CNN early edition", for which the SNR is rather low (presence of music, telephonic calls, noises…etc.) and where the training sub-set is different from the testing one.

In order to evaluate the different techniques described above, several experiments of speaker discrimination are done on the previous database, each experiment concerns one particular method and the corresponding results are represented on Figs. 3 and 4.

The figures represent the ROC curve for the two classifiers: NN and statistical measure. We notice that the NN gives an EER of 9.25 % while the EER given by the statistical measure is 11.75 %. The NN looks better than the statistical method in the middle area, whereas at the borders of the ROC curve, the statistical measure looks better.

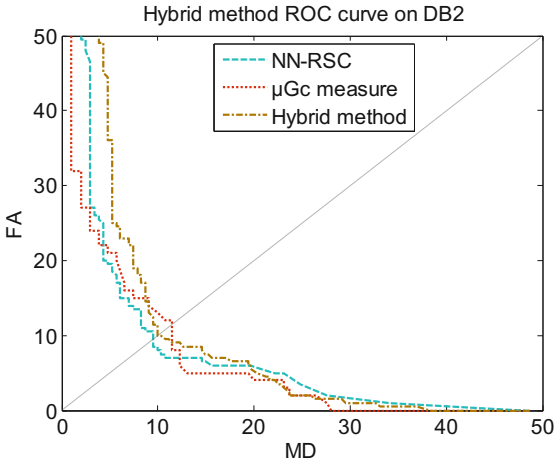

Fig. 3. Speaker discrimination –Hybrid Method-

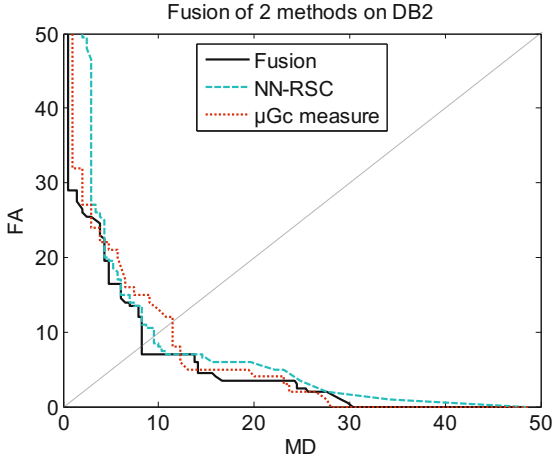

Fig. 4. Speaker discrimination –Fusion NN/μ_{Gc}-

In the other hand, this EER is about 9.95 % when we use the hybrid method (Fig. 3), which means an improvement of 1.8 % with respect to the statistical measure score and a degradation of 0.7 % with respect to the NN.

Results of fusion between the two classifiers: NN and μ_{Gc} are shown in Fig. 4, where we can notice that the fusion gives an EER better than the EER given by each method alone. The fusion EER is only 7.88 % (Table 1) which shows that the fusion is useful. The overall results are summarized in Table 1.

Table 1. Equal Error Rates for different methods.

Classifier / Method	EER %
Statistical measure	11.75
NN-RSC	9.25
Hybrid method	9.95
Fusion (*weighted sum*): NN- μ_{Gc}	7.88

4 Conclusion

Speaker discrimination consists in checking if two different speech segments are uttered by the same speaker or by two different speakers. In order to deal with this problem, several techniques are developed. In this paper, we are interested in four methods, namely: MLP based method, statistical measure based method (μ_{Gc}), hybrid method (MLP-μ_{Gc}) and even a fusion (at score level) based method for the task of discrimination. All those methods are evaluated on a sub-set extracted from Hub4 Broadcast News database and the different scores obtained by each method are represented in a way of ROC curves.

Results allow us to do some comparisons between those four methods according to their corresponding EER.

In one hand, we notice that The NN EER is better than the μ_{Gc} one, which confirms once again the high discriminative capacity of neural networks [12]. In the other hand, the hybrid method resulting from the mixture of the NN and the statistical method has a medium EER of 9.95. The fourth method tested here is the fusion technique carried out with the two basic classifiers. This technique combines the different scores obtained by each method, with a specific weighting coefficients of confidence. This fusion has highly improved the precision of speaker discrimination with an EER of 7.88 % (best score obtained).

In the overall, this research work has shown the difficulties encountred in speaker discrimination, the high discriminative properties of NNs and the relevance of the fusion technique. For future works, we hope to expand our experiments to other fusion techniques used for the same task.

References

1. Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J.: The Development of the 1996 HTK broadcast news transcription system. In: Workshop DARPA Speech Recognition, pp. 97−99 (1997)
2. Motlicek, P., Dey, S., Madikeri, S., Burget, L.: Employment of subspace gaussian mixture models in speaker recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, pp. 4445–4449, 19–24 April 2015
3. Richardson, F., Reynolds, D., Dehak, N.: Deep neural network approaches to speaker and language recognition. IEEE Sig. Process. Lett. **22**(10), 1671–1675 (2015)

4. Ouamour, S., Sayoud, H.: Speaker detection on telephone calls using fusion between SVMs and statistical measures. In: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Beijing, China, 10–12 October 2013
5. Alam, M.M., Uddin, M.S., Uddin, M.N.: Text dependent speaker identification using hidden Markov model and mel frequency cepstrum coefficient. Int. J. Comput. Appl. **104**(14), 33–37 (2014)
6. Lee, H.S., Tsoi, A.C.: Application of multi-layer perceptron in estimating speech / noise characteristics for speech recognition in noisy environment. Speech Commun. **17**(1–2), 59–76 (1995)
7. Sayoud, H., Ouamour, S., Boudraa, M.: 'ASTRA' an automatic speaker tracking system based on SOSM measures and an interlaced indexation. acta. Acustica **89**(4), 702–710 (2003)
8. Sayoud, H., Ouamour, S.: Reconnaissance automatique du locuteur en milieu bruité. In: JEP 2000 Conference, Aussois Juin, pp. 345−348 (2000)
9. Ouamour, S., Sayoud, H.: Looking for the best spectral resolution in automatic speaker recognition. In: 3rd IEEE-GCC Conference, Manama Bahrain, 19–22 March (2006)
10. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L.: Second-order statistical measures for text-independent broadcaster identification. Speech Commun. **17**(1–2), 177–192 (1995)
11. Bonastre, F., Besacier, L.: Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur. Actes du 4ème Congrès Français d'Acou., Marseille pp. 357–360, 14–18 Apr 1997
12. Bennani, Y.: Approches connexionnistes pour la reconnaissance du locuteur: modélisation et identification. Ph. D. thesis, Université Paris XI (1992)
13. Sayoud, H.: Automatic speaker recognition using neural approaches. Ph. D. thesis, USTHB University, Algiers (2003)
14. Ouamour, S., Guerti, M., Sayoud, H.: A new relativistic vision in speaker discrimination. Can. Acoust. J. **36**(4), 24–34 (2008). Publisher: Canadian Acoustics Association, Canada
15. Dasarathy, B.V.: Decision Fusion. IEEE Computer Society Press, Los Alamitos (1994)
16. Kitler, J.: Multiple classifier systems in decision-level fusion of multimodal biometric experts. 1st BioSecure residential workshop, Paris, France 1–26 August (2005)