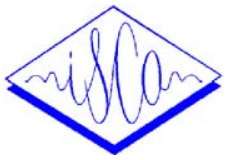

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ON THE USE OF SUPRA MODEL INFORMATION FROM MULTIPLE CLASSIFIERS FOR ROBUST SPEAKER IDENTIFICATION

Hakan Altınçay and Mübeccel Demirekler

Speech Processing Laboratory
Department of Electrical and Electronics Engineering
Middle East Technical University, P.K. 06531, Ankara, TURKEY
E-mail: {auto, demirek}@metu.edu.tr

10.21437/Eurospeech.1999-237

ABSTRACT

In this paper, we propose a text-independent speaker identification (SI) scheme under uncertainty. In this scheme, extraction of *supra model* information about probability distributions in the feature space is proposed. *Supra modeling* is a *model* clustering technique which groups the speaker models into model sets where the speakers in these sets have similar properties. The scheme uses the Dempster-Shafer (D-S) theory of evidence to combine the model sets of two classifiers which are thought to provide complementary information about the speaker identity. A dependency analysis of classifiers to be combined is presented and it is shown to be effective in avoiding wrong decisions. Experimental results of the classifier combination system is given at the end of the paper.

1. INTRODUCTION

As a pattern recognition task, speaker identification is a difficult problem because, i) it involves large amount of noise, ii) insufficient training data and iii) high dimensional feature vectors.

i) Practical applications of speaker identification involves the use of telephone channels. So robust classifiers should be developed to deal with the variabilities produced by different channel conditions. In order to deal with the variability in the channel characteristics, *homomorphic deconvolution* which is a translation from frequency domain to a logarithmic domain (also known as *cepstral* domain) is used where the speech signal is separated from the linear time-invariant (LTI) channel filter [1]. The channel component comes out as an additive part which is assumed to be a constant vector during the complete session and it is subtracted from the signal. This is also known as cepstral mean subtraction (CMS) [1]. The drawback of this approach is that some valuable information about the speaker characteristic is also lost after CMS. Although it is assumed constant in the CMS method, the characteristic of telephone channels may have variabilities within a session and this results in shifts of the feature vectors in the vector space because of the additive effect of the channel in the cepstral domain. There are also some other sources of channel noise [1].

ii) The performance of speaker identification systems is generally tested on databases. The amount of training data in the database, the speech recording conditions and the length of the test sessions vary among different databases. Even within a database, the length of training sessions vary among speakers. Some speakers have sufficient amount of training data where the length of training sessions for some other speakers may be short. Consequently, the models of some speakers may be poor compared to the models of other speakers. If this is the case, the performance of the identification system may be high for speakers with sufficient training data but very low for those speakers with insufficient training data.

This paper presents a speaker identification (SI) scheme under uncertainty. The proposed method deals with the problems given in (i) and (ii). In this scheme, extraction of *supra model* information about probability distributions in the feature space is proposed. *Supra modeling* is a *model* clustering technique to group the speaker models into model sets. The speakers in these sets have similar properties. For instance, the set named as *Sure Set* consists of the speakers whose models are trained with sufficient training data and the classifier has no difficulty in identifying them. *Bad Set* consists of the speakers whose models are poor and the classifier has severe problems in identifying these speakers. Our experiments have shown that instead of characterizing the speakers only with their probability distributions (i.e. selecting the speaker whose model has maximum likelihood), considering their likelihood values together with the ranking of the *neighbor* speakers provides robustness against acoustical channel mismatch. The word *neighbor* will be frequently used in this paper. For a given speaker, the set of speakers whose models are close to that speaker in the feature space are named as the *neighbors* of that speaker and they are grouped as the *Neighbor Set* of the speaker under investigation.

Classifier combination is shown to be effective in developing robust classification systems [2]. In general, there exists a number of classifiers developed using different features and classification schemes. Each of these classifiers may provide a certain degree of success but none of them achieve the expected performance in practical applications. However, since different feature sets represent the pattern classes from different viewpoints and similarly different classification techniques assume different probabilistic models about the pattern classes, for different classifiers the pattern classes which are misclassified may not necessarily overlap and consequently will provide complementary information about the pattern classes. Many different combination schemes are proposed so far which are in the form of rule-based formulations, Bayesian formalism or Dempster-Shafer (D-S) theory of evidence etc. [3], [4]. In this paper, a combination scheme is proposed which is used to combine the *supra models* of multiple classifiers. The combination is done on the basis of D-S theory of evidence which is highly suitable to handle uncertainty.

2. CLASSIFIERS AND DATABASE

In this study, two similar classifiers are developed. The difference is in the feature vectors they use. For both of the classifiers, 12 MFCC and 12 Δ-MFCC coefficients are computed which are concatenated to form a 24 element feature vector per frame. For *classifier #1*, cepstral mean subtraction (CMS) is applied to the features to minimize the channel variation effects but it not applied to *classifier #2* since cepstral means also contain speaker information [1]. For feature extraction, speech signals are blocked into frames of length 20 ms with 10 ms overlapping for the short-time spectral analysis. Then the speech signals are automatically segmented into 4 broad sound classes as voiced, unvoiced, transition and silence. A GMM is trained for

each 4 different speech parts i.e. for each speaker, a GMM is trained using only voiced segments, another GMM for unvoiced and one for transitional regions [5]. For silence regions, a single GMM is trained which is common to all speakers. During testing, the output of the model giving the largest likelihood value is used. The experiments are conducted for the first 30 male speakers of the POLYCOST database. This database consists of text-independent training sessions where the speakers talked in their native languages. A utterance from each of the first two sessions of the database that contain free text speech are used for training, a speech utterance from session three is used for validation and sessions starting from 5 are used for testing. The average length of training sessions is approximately 20 seconds. For 30 speakers, there are 173 test sessions. All sessions are recorded on telephone lines and sampled at 8kHz.

3. SUPRA MODEL SETS

In this section, a brief summary of the supra model sets extracted from each classifier is given. The information listed below is extracted by using a validation utterance for each speaker. Table 1 and Table 2 are examples that show the most likely 5 speakers when *classifier #1* and *classifier #2* are tested by the validation utterances. Note from Table 2 that, for speaker S_2, instead of 5 speakers, there are 3 speakers in its most likely set. The reason for this is that the likelihood of the speaker in the fourth position is very close to zero and it is not taken into account.

Neighbor Set, S_i^N***:*** The ordered set of most likely N speakers obtained by testing the classifier with the validation data of speaker S_i. Speaker S_i may or may not be in this set. If not, learning the neighbors of S_i provides us some *a priori* information about the model of speaker S_i. For example, $S_1^N = \{S_{25}, S_{24}, S_{17}, S_{23}, S_{10}\}$ for *classifier #1*.

Likelihood Set, S_i^L***:*** Speaker S_j is included in the *Likelihood Set* of the speaker S_i, if S_i is in the *Neighbor Set* of S_j. Furthermore, S_i^L is enlarged to include the speaker S_i when S_i is not in the *Neighbor Set* of itself. As an example from Table 1, $S_3^L = \{S_3, \ldots, S_{13}, \ldots, S_{30}\}$.

Bad Set, S_B***:*** This is the set of speakers for which the most likely speaker, i.e. the first element of S_i^N, is not S_i when the classifier is tested by the validation data, or it is the most likely speaker but with a likelihood ratio $\eta < \tau_v$. τ_v is a predetermined threshold and η is defined as $\eta = \frac{L_1}{L_2}$ where L_1 and L_2 are respectively the likelihood values of the most likely and the second most likely speakers that are obtained from model testing. The speakers included in the *Bad Set* are difficult to be correctly identified. Hence, the speakers in this set give us information about the weaknesses of the classifier. Using Table 1, the *Bad Set* is obtained as $S_B = \{S_1, \ldots, S_{13}, \ldots\}$. From the classification performance point of view, the classifier with smaller *Bad Set* cardinality is more powerful compared to a classifier with larger *Bad Set* cardinality.

Sure Set, S_{sure}***:*** The set of speakers which satisfy $|S_i^L| = 1$. The speakers in this set give us information about the strength of the classifier. From the identification performance point of view, the classifier with larger *Sure Set* cardinality is more powerful compared to another classifier with smaller *Sure Set* cardinality. From Table 1, $S_{sure} = \{S_2, \ldots\}$.

The information listed below is extracted during testing of an unknown speaker.

Decision Set, S_D***:*** The set of most likely D speakers resulting during testing the classifier with an unknown utterance.

Bad Set	Speaker tested	Most likely 5 speakers
▷	1	$S_{25}, S_{24}, S_{17}, S_{23}, S_{10}$
	2	$S_2, S_9, S_{19}, S_{27}, S_{10}$
	3	$S_3, S_{20}, S_{16}, S_{25}, S_{17}$
⋮	⋮	⋮
▷	13	$S_{18}, S_{13}, S_{20}, S_3, S_1$
⋮	⋮	⋮
	30	$S_{30}, S_{15}, S_3, S_{20}, S_{17}$

Table 1: Most likely 5 speakers with corresponding validation data for *classifier #1*.

Bad Set	Speaker tested	Most Likely 5 Speakers
▷	1	$S_{24}, S_6, S_{17}, S_{12}, S_{27}$
	2	S_2, S_{16}, S_{19}
	3	$S_3, S_{25}, S_{27}, S_{17}, S_8$
⋮	⋮	⋮
▷	10	$S_{12}, S_6, S_{23}, S_{24}, S_{10}$
⋮	⋮	⋮
▷	30	$S_{13}, S_{30}, S_{18}, S_{23}, S_{17}$

Table 2: Most likely 5 speakers with corresponding validation data for *classifier #2*.

Selected Decision Set, S_d***:*** It is defined in Table 5 for different cases.

In order to discriminate between the supra model sets of two classifiers, the number of the classifier is used as a subscript in the labels of the sets. For example $S_{1,i}^N$ denotes the *Neighbor Set* of speaker S_i in *classifier #1*.

4. DEPENDENCY OF CLASSIFIER DECISIONS

The dependency of classifier decisions is a vital problem in decision fusion. This problem comes into picture especially when the classifiers agree on a wrong speaker. In order to avoid this, we should be able to discriminate between the cases when both of the classifiers are in agreement and the decisions are correct and the case when the decision they agree is actually wrong and the agreement is due to dependency of classifiers for that particular decision. Also, if we somehow decide that the speaker on which both classifiers agree may be a wrong one, we should be able to select a set of possible candidates which include the correct speaker.

4.1. Dependency Definition

For the two classifier decision fusion problem, a common decision on a speaker S_i is defined as a dependent decision if and only if

1. $\exists S_j \neq S_i$, $S_i \in S_{1,j}^N$, $S_i \in S_{2,j}^N$ and
2. $S_j \in S_{1,B}$ and $S_j \in S_{2,B}$

Then, $S_{i,dep}$ is defined as the set $S_{i,dep} = \{S_i, S_j, \ldots\}$. Note that each speaker S_i is always an element of $S_{i,dep}$. If there does not exist any S_j satisfying both of the above conditions, decision S_i is said to be an independent decision and its dependent set is defined as $S_{i,dep} = \{S_i\}$.

If S_i is a dependent decision and both classifiers agree on S_i, we do not make our joint decision on that particular speaker

classifier #1	*classifier #2*
$m\{S^L_{1,i}\} = \alpha_1$	$m\{S^L_{2,j}\} = \alpha_2$
$m\{S_{1,D}\} = \beta_1$	$m\{S_{2,D}\} = \beta_2$
$m\{\Theta\} = \gamma_1$	$m\{\Theta\} = \gamma_2$

Table 3: Source of evidences and their bpa's.

S_i directly, but consider the set of speakers $S_{i,dep}$. From Table 1 and Table 2, we find that decision on speaker S_{24} is a dependent decision and when both classifiers come to the same decision on speaker S_{24}, we should also consider S_1 as a candidate for joint decision since $S_{24,dep} = \{S_1, S_{24}, \ldots\}$. In the next section, the basic probability assignment (bpa) [3] for the supra model sets is described.

5. SUPRA MODEL SETS AND THEIR BPA ASSIGNMENTS

Consider the case where two classifiers are used. For each classifier, *Neighbor Sets*, $S^N_{m,i}$, and *Likelihood Sets*, $S^L_{m,i}$, are calculated for all speakers where m shows the classifier, and the sets $S_{m,B}$ and $S_{m,sure}$ are calculated for both classifiers during the training phase. The raw output of each classifier is its *Decision Set*, $S_{m,D}$. The frame of discernment, Θ, is the set of all speakers in the closed set speaker identification problem. Without any supra modeling, assume that the correct identification rates of the classifiers are r_1 and r_2 respectively. We define the bpa of the evidences as shown in Table 3. Bel(Θ)=1 implies that $\alpha_m + \beta_m + \gamma_m = 1$ for $m = 1, 2$.

α_m, β_m and γ_m, $m = 1, 2$ are design parameters and are selected according to the rules stated below:

- If $r_1 > r_2$, i.e. the performance of *classifier #1* is better compared to *classifier #2*, then $\alpha_1 > \alpha_2$ and $\beta_1 > \beta_2$.
- It seems reasonable to select the values of α's and β's as $\alpha_m + \beta_m > r_m$ since r_m gives the probability that the correct speaker will be the most likely speaker.
- Under the assumption that the most likely N speakers obtained from the validation data is similar to the most likely N speakers when the test data is used, it is reasonable to select $\alpha_m > \beta_m$. When the most likely N speakers do not change, the *Likelihood Set* will always include the correct speaker while the correct speaker may not exist in most likely N speakers.

For our speaker identification experiment, we chose $\alpha_m + \beta_m = r_m$ for $m = 1, 2$. For determining the values of α_m and β_m, a new variable k_m is defined for the *classifier # m* which is calculated by using the validation data and is related with the reliability of *Neighbor Sets*. It is defined as the number of S^N_i sets which include speaker S_i. The reliability increases as the number of S^N_i sets including the speaker S_i increase. Then the values of α_m and β_m are related by

$$\frac{k_m}{R}\alpha_m = \beta_m \tag{1}$$

where R is the total number of speakers in our closed set. Note that we also had the equation $\alpha_m + \beta_m = r_m$. From the simultaneous solution of these equations, for $m = 1, 2$, we obtain

$$\alpha_m = \frac{R.r_m}{R + k_m} \quad and \quad \beta_m = r_m - \alpha_m. \tag{2}$$

Θ	S^L_i				
S_f					
S_e	$S_e \cap S^L_i$				$S_e \cap \Theta$
S^L_j	$S^L_j \cap S^L_i$				
$\oplus$	S^L_i	S_a	S_b	S_c	Θ

Table 4: Application of D-S rule of combination

5.1. Refinement Operation

We do not have further information to discriminate among the speakers in the set $S^L_{m,i}$. However $S_{m,D}$ can be partitioned into smaller sets by using a similarity measure between $S_{m,D}$ and the neighbor sets of each speaker in $S_{m,D}$. It is first partitioned into the sets $S_{m,d}$ and $\{S_{m,D}\} \setminus \{S_{m,d}\}$ where $S_{m,d}$ is defined in Table 5 for different cases. β_m is assigned only to $S_{m,d}$ and it is distributed to each element of $S_{m,d}$ according to the following expression:

$$m\{S_k\} = \frac{\kappa |S_{m,D} \cap S^N_{m,k}|}{|S^N_{m,k}|} \quad \forall S_k \in S_{m,d} \tag{3}$$

where κ is selected such that

$$\sum_{S_k \in S_{m,d}} m\{S_k\} = \beta_m. \tag{4}$$

After assigning the bpa values to both classifiers, we apply the D-S rule of combination.

6. APPLICATION OF DEMPSTER-SHAFER EVIDENCE COMBINATION

Table 4 shows the resultant operations of evidence combination. From this point on, it is going to be assumed that S_i and S_j are the most likely speakers of *classifier #1* and *classifier #2* respectively. Assume that $S_{1,d} = \{S_a, S_b, S_c\}$ and $S_{2,d} = \{S_e, S_f\}$. The sets resulting from the intersections (only some of them are shown on the table for brevity) will have bpa values which are the products of bpa's of the intersected sets, i.e. $m\{S^L_i \cap S_f\} = m\{S^L_i\} m\{S_f\}$. These bpa's are then normalized by distributing the bpa's of empty sets among the nonempty sets and setting the bpa's of empty sets to zero. After combination, let the set $\{A_1, A_2, A_3, \ldots, A_K\}$ contain all possible nonempty sets. The betting probability of a speaker S_n is calculated as [6];

$$P\{S_n\} = \sum_{i=1}^{K} m(A_i) \frac{|S_n \cap A_i|}{|A_i|}. \tag{5}$$

Then the decision is made on the speaker with maximum betting probability.

D-S evidence combination algorithm is applied in a rule-based manner. The bpa values are assigned depending upon the classifier decisions. The dependency problem of classifier decisions is solved in this way. The rule-based application of D-S formalism is summarized in Table 5. Starting from *Case 1*, the case that matches the behavior of the classifiers is used to make the joint decision.

In cases 1 and 2, a decision on a speaker in the *Sure Set* is made directly. This means that the corresponding classifier is capable of surely discriminating the speakers in its *Sure Set* from others. In such a case, it is assumed that there is no uncertainty in the decision.

In case 3, the classifiers agree on the same speaker and the dependency set of this particular speaker does not include any other speaker, i.e. $S_{i,dep} = \{S_i\}$.

Case 1	$S_i \in S_{1,sure}$
	$S_f = S_i$
Case 2	$S_j \in S_{2,sure}$
	$S_f = S_j$
Case 3	$S_i = S_j$
	$S_{i,dep} = \{S_i\}$
	$S_f = S_j$
Case 4	$S_i = S_j$
	$\exists S_k \neq S_i, S_k \in S_{i,dep}$
bpa	$S_{m,d} = S_{i,dep} \cap S_{m,D}$
	$m\{S_{m,dep}\} = \alpha_m$
	$m\{S_{m,d}\} = \beta_m$
	$m\{\theta\} = 1.0 - \alpha_m - \beta_m$
	Refine $S_{m,d}$ into its singletons, i.e
	$\forall S_k \in S_{m,d},\ m\{S_k\} = \frac{\kappa \lvert S_{m,D} \cap S^N_{m,k} \rvert}{\lvert S^N_{m,k} \rvert}$
	Apply $m_1 \oplus m_2 \Longrightarrow S_f$=?
Case 5	$S_i \neq S_j$
	$S_i \notin S_{1,B}$
	$S_j \notin S_{2,B}$
bpa	$S_{m,d} = S_{m,D} \cap S_{m,B}$
	$m\{S^L_{m,i}\} = \alpha_m$
	$m\{S_{m,d}\} = \beta_m$
	$m\{\theta\} = 1.0 - \alpha_m - \beta_m$
	Refine $S_{m,d}$ into its singletons, i.e
	$\forall S_k \in S_{m,d},\ m\{S_k\} = \frac{\kappa \lvert S_{m,D} \cap S^N_{m,k} \rvert}{\lvert S^N_{m,k} \rvert}$
	Apply $m_1 \oplus m_2 \Longrightarrow S_f$=?
Case 6	$S_i \neq S_j$
	$S_i \in S_{1,B}$
	$S_j \notin S_{2,B}$
bpa	$S_{m,d} = S_{m,D} \cap S_{m,B}$
	$m\{S^L_{m,i}\} = \alpha_m$
	$m\{S_{m,d}\} = \beta_m$
	$m\{\theta\} = 1.0 - \alpha_m - \beta_m$
	Refine $S_{m,d}$ into its singletons, i.e
	for m=1:
	$\forall S_k \neq S_i \in S_{m,d},\ m\{S_k\} = \frac{\kappa \lvert S_{m,D} \cap S^N_{m,k} \rvert}{\lvert S^N_{m,k} \rvert}$
	for $S_k = S_i\ \ m\{S_k\} = \frac{\gamma . \kappa \lvert S_{m,D} \cap S^N_{m,k} \rvert}{\lvert S^N_{m,k} \rvert}$
	for m=2:
	$\forall S_k \in S_{m,d},\ m\{S_k\} = \frac{\kappa \lvert S_{m,D} \cap S^N_{m,k} \rvert}{\lvert S^N_{m,k} \rvert}$
	Apply $m_1 \oplus m_2 \Longrightarrow S_f$=?

Table 5: Bpa assignments for Dempster's rule of combination.

Classifier	*Identification Rate*
Classifier #1	80.9%
Classifier #2	77.3%
D-S Based System	91.9%

Table 6: Comparison of results

Case 4 deals with the problem of classifier dependency. In this case, possible dependent speakers that are learned from the validation data are assigned nonzero bpa's and from the decision sets, only those speakers are refined for the decision combination process. Our experiments have shown that, in the cases where the decisions of two classifiers are identical (both are S_i) but are wrong decisions, $S_{i,dep}$ always includes the correct speaker.

Classifiers are not capable of identifying speakers in their *Bad Sets*. In the refinement operation, only the speakers that are in the *Bad Set* of the classifier are refined (as done in case 5, $S_{m,d} = S_{m,D} \cap S_{m,B}$). This forces the decision combination operation to give a priority to bad speakers.

Case 6, deals with the case that the most likely speaker of a classifier is in its *Bad Set*. In case 5, we forced decisions on bad speakers by selecting the bad speakers from the decision set during the refinement operation. But in this case, more than this, we force the decision on S_i, which is a bad speaker, by further scaling the bpa value to that speaker before normalizing the bpa's of refined speakers. Note from equation (3) that the variable κ is used for this normalization so that (4) is satisfied. The scaling by the factor γ is applied before this normalization.

7. EXPERIMENTAL RESULTS

For $N = 5$, $D = 10$, $\tau_1 = 10^5$, $\tau_2 = 10^{25}$, $R = 30$ male speakers and $\gamma = 3.0$, the identification performances of individual classifiers and the resulting performance after decision combination are shown in Table 6. As seen from the table, a considerable improvement is achieved by the combination of the decisions of the classifiers.

8. CONCLUSION

In this paper, some information sources based on supra model sets are presented. These sources of information are shown to be effective for speaker identification. Combination of outputs of two classifiers was another main subject of this study. This is done in a rule-based manner. The concept of independence of decisions developed in this study is also considered in obtaining a joint decision. The combined system, which is based on the D-S theory of evidence, surpassed the performance of individual classifiers by a rate of 11.0% and 14.6% respectively for *classifier #1* and *classifier #2*.

9. REFERENCES

[1] H. Gish and M. Schmidt: "Text-independent speaker identification.", *IEEE Signal Processing Magazine*, pp. 18-32, Oct., 1996.

[2] L. Xu, A. Krzyzak and C. Y. Suen: "Methods of combining multiple classifiers and their applications to handwriting recognition", IEEE Trans. on Systems, Man. and Cybernetics, vol. 22, pp. 418-435, 1992.

[3] G. Shafer: "A Mathematical Theory of Evidence", Princeton University Press, 1976.

[4] M. Demirekler and A. Saranlı: "A study on improving decisions in closed set speaker identification.", *in IEEE-ICASSP Proceedings*, pp. 1127-1130, 1997.

[5] D. A. Reynolds and R. C. Rose: "Robust text-independent speaker identification using Gaussian mixture models.", *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 72-83, 1995.

[6] P. Smets and T. Kennes: "The transferrable belief model.", *Artificial Intelligence*, vol. 66, pp. 191-234, 1994.