USE OF MODEL CONFUSION LEARNING FOR SPEAKER IDENTIFICATION A RULE-BASED APPROACH

Hakan Altınçay and Mübeccel Demirekler

Speech Processing Laboratory
Department of Electrical and Electronics Engineering
Middle East Technical University, P.K. 06531, Ankara,Turkey
email: {auto,demirek}@rorqual.cc.metu.edu.tr

ABSTRACT

This paper presents a multiple classifier system for text-independent speaker identification (SI). For the speaker identification problem, several different classifiers can be developed, each having strengths and weaknesses compared to the others. When the strengths and weaknesses of the individual classifiers do not overlap, i.e. a speaker which is misclassified by one classifier is correctly classified by some others, robust classification systems can be developed with the use of multiple classifiers. The studies in multiple classifier systems mainly concentrate on reliable methods of extracting contextual information (i.e. strengths and weaknesses) about the classifiers and the methods of combining these classifiers. In this paper, a method is proposed for the extraction of contextual information about the classifiers and a rule based approach is developed for the combination of the information from different classifiers.

1. INTRODUCTION

The performance of a single classifier speaker identification system (SI) may be degraded because of insufficient training data and acoustical channel mismatch between the training and test sessions. When the training data is insufficient, the estimated probability distributions do not correctly characterize to corresponding speakers and consequently these speakers are not correctly identified. In the cases where noisy telephone channels are used, the estimated speaker models may be shifted in the feature space and consequently they may highly overlap with the models of other speakers. In order to deal with these problems, a method of extracting contextual information (CI) about the classifiers is proposed. Using the proposed approach, it is intended to learn

- The speakers whose models are well separated from the models of other speakers in the feature space
- The speakers whose models highly overlap with each other in the feature space
- The speakers whose models do not correctly characterize the corresponding speaker

The approach proposed groups the speakers under some sets. 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 probably wrong 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. Suppose that an utterance of an unknown speaker is tested and the speaker labeled as S_i came out to be the most likely speaker. Then in order to justify the decision on that speaker, it is expected that the *neighbor* speakers of this speaker are more likely compared to the other speakers. From the *Neighbor Set*s, the speaker sets named as *Likelihood Set*s are derived. The *Likelihood Set* of speaker S_i is the set of all speakers which involve speaker S_i in their *Neighbor Set*s. When the speaker S_i comes out to be the most likely speaker after testing the utterance of an unknown speaker, in order to avoid classification errors, all speakers in the *Likelihood Set* of S_i are considered as possible candidates for the decision. It should noted that each speaker is an element of its *Likelihood Set*.

The philosophy behind information combination using multiple classifiers is to make use of the advantage of each classifier [1]. If the combination is not done carefully, the correct information coming from a classifier may be destroyed by the misleading information coming from other classifiers. Hence, it is very important to be able to decide

the cases where a classifier performs better compared to the others. This can only be achieved by extracting some contextual information about the classifiers [2]. The method proposed in this paper is used for this purpose and then a rule based combination scheme is developed which is used to combine the information coming from two classifiers.

2. CLASSIFIERS AND DATABASE

For the combination problem, two classifiers are developed. For both of the classifiers, 12 Mel frequency cepstral coefficients, 12-MFCC and 12 Δ-MFCC coefficients are computed which are concatenated to form a 24 element feature vector per frame. For the first classifier, cepstral mean subtraction (CMS) is applied to the features but not for the second classifier. Experiments have shown that CMS, which is used to solve channel mismatch problem may in some cases remove the speaker identity [3]. 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 four sets, i.e. for speaker S_i, a GMM is trained using only voiced segments, another GMM for unvoiced and one for transitional regions [4]. For silence regions, a single GMM is trained which is common to all speakers by using the silence regions from the training data of all speakers. During testing, the output of the model giving the largest likelihood value is used.

Experiments are conducted for the first 30 male speakers of the POLYCOST database [5]. This database consists of text-independent training sessions where the speakers talked in their native languages. There are speakers from 14 different countries. For each speaker, a utterance from each of the first two sessions (mot02 files) are used for training and a utterance from session 3 (mot01 file) is used for validation. The mot01 files from the sessions starting from 5 are used for testing. The average length of the training sessions is around 20 seconds including the silence regions. For 30 speakers, there are 173 test sessions. All sessions are recorded on telephone lines and sampled at 8kHz.

3. CONTEXTUAL INFORMATION

In this section, a brief summary of the information extracted from each classifier is given. The information listed below is extracted by using a validation utterance for each speaker. Table 1 is an example that shows the most likely 5 speakers when *classifier #1* is tested by the validation utterances of the corresponding speakers.

3.1. *Neighbor Set,* S_i^N

This is the ordered set of most likely N speakers obtained by testing the classifier with the validation data of speaker S_i. The speaker S_i may or may not be in this set. If not, this means that the training data has missing sound classes which exist in the test utterance or the training data is so short that the estimated model is not the correct probability distribution of the corresponding speaker. For example. using Table 1, for speaker S_1, $S_1^N = \{S_{25}, S_{24}, S_{17}, S_{23}, S_{10}\}$ for *classifier #1*. During testing, if S_1 comes out as the most likely speaker of *classifiers #1*, it will be selected as the winner only if the neighbors of this speaker are placed in the top ranks. Otherwise, we will be suspicious about the decision on S_1 and the developed algorithm will consider some other criteria to make a decision. In this study, we selected $N = 5$.

3.2. *Likelihood Set,* S_i^L

A 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. The *Likelihood Set* S_i^L denotes the set of speakers which are treated as equally likely to be the correct speaker when the speaker S_i turns out to be the most likely speaker. As an example from Table 1, for speaker S_3, $S_3^L = \{S_3, \ldots, S_{13}, \ldots, S_{30}\}$. This set has a very important function in the developed algorithm. As described in section 1, when the estimated model of a speaker does not correctly characterize the corresponding speaker, he/she will not come out to be the most likely speaker when his/her utterance is tested. In other words, this speaker is confused with some other speakers. Learning the confused speakers forms the *Likelihood Set* by which information about the correct speaker is not lost.

3.3. *Bad Set,* S_B

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

$$\eta = \frac{L_1}{L_2} \tag{1}$$

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* can easily

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*.

be 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 larger *Bad Set* cardinality.

3.4. *Sure Set,* S_{sure}

The set of speakers satisfying $|S_i^L| = 1$. In other words, the speakers which are not included in the *Neighbor Set* of any other speaker. The classifier does not have any difficulty in identifying these speakers. The speakers in this set give us information about the strength of the classifier. From the classification performance point of view, the classifier with larger *Sure Set* cardinality is more powerful compared to a classifier smaller *Sure Set* cardinality. From Table 1, $S_{sure} = \{S_2, \ldots\}$.

3.5. *Decision Set,* S_D

The ordered set of most likely D speakers resulting from testing the classifier with the speech data of an unknown speaker. Note that the *Neighbor Set*s are identical to the *Decision Set*s when validation sessions are used.

4. SINGLE CLASSIFIER CASE

In this section, the algorithm to use the contextual information for improving the identification performance of *classifier #1* is given. In the algorithm, it is assumed that the first element of the decision set S_D is the speaker S_i.

4.1. Algorithm 1

Step 1 If $S_i \in S_{sure}$ or $\eta > \tau_t$, then select S_i, as the winner, else goto Step 2.

Step 2 If $S_i \in S_B$ then this speaker is the winner, else goto Step 3.

Step 3 Give a second chance to the most likely speaker (S_i in this case). If $|S_D \cap S_i^N| \geq \alpha$ then the decision is on S_i. The reason for this is that the *neighbors* of the most likely speaker are same both for the validation data and the current test data. Otherwise goto Step 4.

Step 4 Find a subset of the decision set S_D, say S_d where $S_d = S_D \cap S_B$. Then for all $S_j \in S_d$, check if $|S_j^N \cap S_D| \geq \gamma$. If only one speaker satisfies this condition, then this speaker is the winner. If more than one speaker satisfies it, namely S_n and S_m, find the set $S_n^N \cap S_m^N$. If there exists a unique speaker in this set, then this is the final winner. Otherwise from the *Neighbor Set*s of S_n and S_m, select the one in which the speaker S_j is more likely. For example if $S_n^N = \{S_i, S_j, S_k, \ldots, S_l\}$ and $S_m^N = \{S_l, S_k, S_j, \ldots, S_p\}$ then the decision is S_n since in the *Neighbor Set* of this speaker, S_j is in second location while it is in the third location for the *Neighbor Set* of S_m. If there are no speakers satisfying $|S_j^N \cap S_D| \geq \gamma$, then goto Step 5.

Step 5 Decrease the thresholds of α and γ by 1 and goto Step 3.

Step 6 End of the algorithm.

In Step 1, it is checked whether the classifier is *sure* about making a decision on the most likely speaker and if so, the decision is made. Being sure about the decision means that the model of the most likely speaker is well separated from the models of the other speakers.

Since the speakers in the set S_B are those that the classifier cannot correctly identify, when a speaker from this set is the most likely, decision is made on that speaker in Step 2.

In Step 3, when the most likely speaker is not in S_B and the system is *not sure* about making a decision on the most likely speaker, we concentrate on the decision set and try to match S_D to the *Neighbor Set* of the most likely speaker. This kind of work means giving a second chance to the most likely speaker. This is done by checking the *neighbors* of the speaker. If this test also fails, then the system is *sure* that the most likely speaker is not the correct speaker and tries to find the correct speaker from the speakers that are in S_D which are also in the set S_B. This means making the decision among the speakers that the system has difficulties in identification.

4.2. Experimental Results

The variables defined above are set to $N = 5$, $D = 6$, $\alpha = 4$, $\gamma = 3$, $\tau_v = 10^{25}$ and for different values of τ_t, the identification performance of the classifier is shown in Table 2.

Value of τ_t	Identification Rate
No CI	80.9% (140/173)
$\tau_t = 10^{32}$	87.9% (152/173)
$\tau_t = 10^{50}$	86.7% (150/173)

Table 2: Identification rates of *classifier #1* with contextual information by using mot01 files.

4.3. Discussions About the Algorithm

In order to increase the system performance, we should decrease the size of set S_B because our experiments show that the classifier which does not use contextual information generally makes identification errors when testing the speakers which are in S_B. Decreasing this set means building better classifiers which make less number of errors. A possible solution to this problem is to extend the identification system to M classifiers where a decreased *Bad Set*, $S_B^{'} = S_{1,B} \cap S_{2,B} \ldots \cap S_{M,B}$ can be obtained. This will decrease the number of possible speakers that the system has difficulty in identification The aim is to find M classifiers such that $S_B^{'} = \emptyset$. However, arbitrary M classifiers with a decreased $S_B^{'}$ are expected to have a higher identification rate compared to only one of those M classifiers.

5. EXTENSION TO 2 CLASSIFIERS

In order to differentiate between the information sets of different classifiers, the subscript m will be used in the sets which will be 1 for *classifier #1* and 2 for *classifier #2*. For each classifier, the *Neighbor Set*, $S_{m,i}^N$, and the *Likelihood Set*, $S_{m,i}^L$ are calculated for all speakers. The *Sure Set* $S_{m,sure}$ and the *Bad Set* $S_{m,B}$ are also calculated for each classifier. During testing, the output of each classifier is its decision set, $S_{m,D}$ where the most likely speaker is the first element of the set. Assume that *classifier #1* gives S_i as the most likely speaker and *classifier #2* gives S_j as the most likely speaker. When the information from two classifiers are combined, the decided speaker is named as the *joint* decision of the combined system. The rule based classifier combination used is formulated as follows.

5.1. Algorithm 2

Step 1 If $i = j$, then the classifiers agree on the most likely speaker so the joint decision is S_i, else goto Step 2.

Step 2 If $S_i \in S_{1,sure}$, then the joint decision is the most likely speaker of *classifier #1*, else if $S_j \in S_{2,sure}$, then the joint decision is the most likely speaker of *classifier #2*. Otherwise goto Step 3.

Step 3 Find the set of most likely speakers, S_L, where $S_L = S_{1,i}^L \cap S_{2,j}^L$. If the intersection is a unique speaker then the joint decision is this speaker. Else if $S_L = \emptyset$ or $|S_L| \geq 2$, then goto Step 4.

Step 4 If only one of the classifiers satisfies $S_{m,k}^L \cap S_{m,B} = \emptyset$, then the most likely speaker of this classifier is the joint decision. If both of the classifiers satisfy this condition, then the winner is the most likely speaker of the classifier which has a higher individual identification rate when no contextual information is used. If non of them satisfies this condition then go to Step 5.

Step 5 Select the classifier m that has the highest individual identification rate when no contextual information is used. For this classifier, find the subset of $S_{m,k}^L$, say S_d, by eliminating speakers that are not in the set $S_{1,B} \cup S_{2,B}$. Then we are left with speakers that are in the *Bad Set* of either classifier. From this point on apply the Algorithm 1 starting from Step 3 which was developed for the single classifier case.

Step 6 End of the algorithm.

In Step 1, we check whether the most likely speakers of both classifiers are the same and if so, we select the common most likely speaker as the final decision and the algorithm stops.

In Step 2, we check whether the most likely speaker of either classifier is in its *Sure Set*, and if so, the most likely speaker of that classifiers is selected as the joint decision. A possible conflict occurs when the most likely decisions of both classifiers are in their *Sure Sets* and the decisions are in conflict. Our experiments show that this is never the case and if it were, a good approach would be the selection of the decision of the classifier with the highest individual identification rate.

In Step 3, it is checked to see whether there is a unique intersection between the *Likelihood Set*s of two classifiers. If not, the intersection set, i.e. the new possibly correct speaker set, becomes smaller by the approval of both classifiers. This means that each classifier decreases the uncertainty of the other classifier by comparing the two most likely speaker sets of both classifiers. A note for this step is that there is an important reason for why we do not use the intersection of the *Decision Sets* as possible candidates for the joint decision. The explanation for this is as follows. Firstly, the correct speaker may not be in the *Decision Sets* (remember that a speaker may or may not be in its *Neighbor Set*) and secondly, for two distinct classifiers using uncorrelated features, the *Decision Sets* may be inconsistent. Two classifiers may confuse a particular speaker with different speakers. In this case, the intersection set may be an empty

Cassifier	SI Rate
$Classifier\ \#1$ First 2 sessions for Models	80.9% (140/173)
$Classifier\ \#2$ First 2 sessions for Models	77.4% (134/173)
$Classifier\ \#1$ First 3 sessions for Models	97.1% (168/173)
$Classifier\ \#2$ First 3 sessions for Models	97.1% (168/173)
$Classifier\ \#1$ & $Classifier\ \#2$ First 2 sessions for Models 3^{rd} session for CI	91.3% (158/173)
$Classifier\ \#1$ & $Classifier\ \#2$ First 3 sessions for Models 3^{rd} Session for CI	100.0% (173/173)

Table 3: Comparison of SI systems with and without using contextual information (CI).

set and combination of two classifiers may remove the valuable information in their *Decision Sets*. This is not case when *Likelihood Set*s are used because the tested speaker exists in the *Likelihood Set*s of all the speakers that are in the *Neighbor Set* of the tested speaker.

In Step 4, we try to concentrate on the classifier whose *Likelihood Set* S_i^L for the most likely speaker S_i does not contain any element from its *Bad Set*. This step is particularly important since when the *Likelihood Set* does not contain any speaker from the *Bad Set*, then the possible risk of error is too low. Our experiments have shown that the classifiers do not in general make identification errors in their decisions on the most likely speaker if the *Likelihood Set* of this speaker does not contain any speaker from the *Bad Set*.

In Step 5, we select one of the classifiers to make the final decision but from the set $S_{m,k}^L$, the elements that are not in the *Bad Set* of either or both of the classifiers are eliminated because if one of these were actually the correct speaker, since they are not elements of the *Bad Set* of either classifier, the classifiers should have reached at a consensus one one of them.

5.2. Experimental Results and Discussions

In our experiments we used $D = 6$ and $N = 5$ for both classifiers, $\tau_v = 10^{25}$ for *classifier #1* and $\tau_v = 10^5$ for *classifier #2*. The identification rate of the combined classifier system is 91.3%. The results of the experiments are given in Table 3. Note that when the validation session is included into the training data of models, the SI rates of both classifiers increase considerably (refer to rows 3 and 4 from the table). The reason for this is that the context (i.e. the words in the recorded text) of all mot01 files is same in all sessions. So when the same context is used for both training and testing, the results are much better than using the contextual information as described in this paper, but training the models by using only 2 sessions (mot02 files where the context of spoken text is completely arbitrary). The last row of the table corresponds to the case where the third record, i.e. validation data which is a mot01 file, is also used during training (as in rows 3 and 4) and contextual information is used. This experiment gave perfect result. This is actually important because as the experimental results show, even with the usage of similar context, the problems arising from insufficient training data and the noisy telephone channels cannot be avoided but the use of the proposed information sets provides robustness against these disorders.

6. CONCLUSION

In this paper, some contextual information sources based on the confusion of models 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. Combination of classifier outputs provided considerable improvement in the identification rate.

7. REFERENCES

[1] T. K. Ho, J. J. Hull, S. N. Sirhari: "Decision Combination in Multiple Classifier Systems.", *IEEE Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66-75, 1994.

[2] I. Bloch: "Information Combination Operators for Data Fusion: A Comparative Review with Classification", *IEEE Trans. on Systems Man. and Cybernetics*, vol. 26, no. 1, pp. 52-67, 1996.

[3] H. Gish and M. Schmidt: "Text-Independent Speaker Identification.", *IEEE Signal Processing Magazine*, pp. 18-32, Oct., 1996.

[4] 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.

[5] H. Melin and J. Lindberg: "Guidelines for Experiments On the POLYCOST Database.", *COST 250 Workshop, Vigo, Spain.*, pp. 59-69, 1996.