A RANK-BASED CLASSIFIER COMBINATION SCHEME FOR SPEAKER IDENTIFICATION BASED ON RANKING STATISTICS

Hakan Altınçay and Mübeccel Demirekler

Speech Processing Laboratory
Department of Electrical and Electronics Engineering
Middle East Technical University, Ankara, Turkey
E-mail: auto@metu.edu.tr, demirek@metu.edu.tr

ABSTRACT

In this paper, we propose a novel rank-based classifier combination scheme under uncertainty for speaker identification (SI). The combination is based on a heuristic method that uses Dempster-Shafer theory of evidence under some conditions. The method is based on the extraction of first and R^{th} level ranking statistics. Using these statistics, the pattern classes are clustered into model sets where the classes in these sets share set specific properties. Some of these model sets are used to reflect the strengths and weaknesses where some others carry class dependent ranking statistics of the corresponding classifier.The experiments conducted on the Polycost database have shown that the proposed approach is more effective compared to some other rank-based combination schemes.

1. INTRODUCTION

For a pattern recognition problem involving large amount of noise, limited amount of training data and high dimensional feature vectors, it is in general difficult to develop a good classifier. For each pattern recognition problem, there exists a number of classifiers which use different features and architectures but none of them alone achieves the expected performance in the practical applications. Researchers have always been actively studying to develop better classifiers. Also, combination of different classifiers have been proposed as an alternative direction to improve the identification performances of classification systems. There have been extensive research in this field and promising results were obtained. Even the trivial combination approaches like averaging or majority voting has been reported to provide improved results compared to the individual classifiers.

Bayesian probability theory, fuzzy sets, possibility theory and Dempster-Shafer evidence theory are most commonly used frameworks for classifier combination. Using these frameworks, the information provided by different classifiers can be combined and a unique joint decision can be obtained [1]. The selection of the framework to be used is, in general, problem dependent but more important than this is the type of the output information extracted from the classifiers. The output information provided by various classification systems is in general in abstract level, in ranking level or in measurement level. For instance, the *highest rank* method which is a rank based combination approach, can be described as follows. When a speech utterance is tested by a classifier, each speaker receives a rank score. Our convention is that the score is largest for the speaker that is ranked on the top. The combined score assigned to a speaker is the maximum of the scores assigned to that speaker by all of the classifiers. The speaker which gets the maximum score is selected as the joint decision. In the *Borda count* combination method, the score assigned to a speaker by each individual classifier is the *sum* of the scores assigned to that speaker by all of the classifiers. The speaker with maximum sum score is selected as the joint decision [2]. *Logistic regression* is a modified version of the Borda count method. In this case, the combined score assigned to a speaker is the weighted linear combination of the individual classifier scores. The weights reflect the relative significance of the classifiers in the combination process. The details of these methods can be found in [2].

In this paper, we applied some well known rank-based combination techniques to SI and then compared their results with the proposed combination scheme. Although SI is a complex pattern classification problem, classifier combination has not yet attracted the interest of the researchers in the field of speaker identification. There are only few studies in the literature where improvements over the individual classifiers are reported. We believe that classifier combination is a promising direction for developing better speaker identification systems.

The proposed method of classifier combination is based on a heuristic approach that uses Dempster-Shafer theory of evidence [3]. We conducted several experiments on the Polycost database. Our experiments have shown that the proposed method performed better compared to the individual classifiers and some other combination techniques.

2. INFORMATION SOURCES

In this section, the information extracted using the raw classifier outputs is described. This information is modeled in terms of sets. These sets are namely *confusion set* Ω_k^j, *neighbor set* $Neig_k^R(t)$, *bad set* B_k and *sure set* S_k. Before giving the definitions of these sets, R^{th} level confusion matrix will be defined.

2.1. R^{th} level confusion matrix

Let N denote the total number of speakers and k denote the classifier number. An NxN confusion matrix is obtained for each classifier by using the cross validation data as explained below. Let T_i denote the total number of tokens that belong to speaker s_i in the cross validation data. Each of these tokens are tested by the classifier e_k and a set of speakers that appears in the top R rank are counted. The amount of these tokens generates the i^{th} row of the R^{th} level confusion matrix, $Conf_k(R)$. For $1 \leq R \leq N$,

$$Conf_k(R) = \begin{bmatrix} n_{11}^{(k,R)} & n_{12}^{(k,R)} & \dots & n_{1N}^{(k,R)} \\ n_{21}^{(k,R)} & n_{22}^{(k,R)} & \dots & n_{2N}^{(k,R)} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N1}^{(k,R)} & n_{N2}^{(k,R)} & \dots & n_{NN}^{(k,R)} \end{bmatrix} \quad (1)$$

where $n_{ij}^{(k,R)}$ is the number of times that the speaker s_j is ranked in the top R rank when the tokens of the speaker s_i are tested by the classifier e_k

2.2. Neighbor sets, $Neig_k^R(t)$

Given a token "t", the R^{th} level neighbor set is defined as the set of speakers that are in the first R rank when tested by the classifier e_k. This set is denoted by $Neig_k^R(t)$. The probability that a test token belongs to the speaker s_i when it is tested by classifier e_k and s_j is in the set $Neig_k^R(t)$, $P(t \in s_i | s_j \in Neig_k^R(t))$, can be approximately obtained from $Conf_k(R)$ as follows:

$$P(t \in s_i | s_j \in Neig_k^R(t)) = \frac{P(t \in s_i, s_j \in Neig_k^R(t))}{P(s_j \in Neig_k^R(t))}$$
$$= \frac{P(s_j \in Neig_k^R(t) | t \in s_i) P(t \in s_i)}{\sum_{l=1}^{N} P(s_j \in Neig_k^R(t) | t \in s_l) P(t \in s_l)} \quad (2)$$

The conditional probabilities that appear in the above expression can be approximated as follows:

$$P(s_j \in Neig_k^R(t) | t \in s_i) = \frac{n_{ij}^{(k,R)}}{T_i} \quad (3)$$

T_i denotes the total number of cross validation tokens that belongs to s_i. Assuming that $T_i = T$ and $P(t \in s_i) = \frac{1}{N} \forall i$, the above expression can be written in terms of the confusion matrix as follows:

$$P(t \in s_i | s_j \in Neig_k^R(t)) = \frac{n_{ij}^{(k,R)}}{\sum_{l=1}^{N} n_{lj}^{(k,R)}} \quad (4)$$

2.3. Confusion sets, Ω_k^j

For the classifier e_k and each speaker s_j, a confusion set Ω_k^j is defined as the set of speakers such that

$$\Omega_k^j = \{ s_i | n_{ij}^{(k,1)} > 0, \;\; s_i \in \Theta \} \quad (5)$$

Ω_k^j includes the speakers s_i for which s_j may come out to be first ranked. As seen from the equation, first rank confusion matrices are used to define Ω_k^j. During testing, given the first ranked speaker, the confusion set of the corresponding speaker provides the list of candidates for determining the correct speaker.

2.4. Bad sets, B_k

For the classifier e_k, the set B_k is defined as the set of speakers such that

$$B_k = \{ s_i | \frac{n_{ii}^{(k,1)}}{T} \leq \tau_B, \;\; s_i \in \Theta \} \quad (6)$$

In other words, the set B_k includes the speakers for which only a portion of the cross validation tokens determined by the threshold τ_B would be correctly classified if the first rank were chosen as the correct speaker.

2.5. Sure sets, S_k

For the classifier e_k, the set S_k is defined as the set of speakers such that

$$S_k = \{ s_i | \left(1 - \frac{n_{ii}^{(k,1)}}{\sum_{j=1}^{N} n_{ji}^{(k,1)}} \right) \leq \tau_S, \;\; s_i \in \Theta \} \quad (7)$$

As seen from the equation, again first rank confusion matrices are used to define S_k. This set gives the list of speakers for which the classifier is efficient in classifying them in the first rank.

3. CALCULATION OF OPTIMAL R, τ_B AND τ_S VALUES

From this point on, we will assume that R is a classifier dependent quantity so we will represent the R value of the classifier e_k by R_k. The criteria that should be taken into account in calculating the optimal values for R_k's are:

1. The cardinality of neighbor sets should be large so that the probability that the correct speaker is in these sets is close to 1.
2. The cardinality of these sets should be small enough so that the speakers in these sets are informative about the correct class.

Let s_c denote the correct speaker that the input token t belongs and s_n denote an arbitrary speaker. Let us define $\delta_{min}(t)$ as follows:

$$\delta_{min}(t) = \min_{\substack{s_n \\ s_n \neq s_c}} (O_c(t) - O_n(t)). \quad (8)$$

where $O_i(t)$ is proportional with the probability that "t" belongs to speaker s_i given the raw data, i.e. $Neig_k^{R_k}(t)$ of classifier e_k. Note that,

$$P(t \in s_n | s_{i_1} \in Neig_k^{R_k}(t), \dots, s_{i_{R_k}} \in Neig_k^{R_k}(t)) =$$
$$\frac{P(s_{i_1} \in Neig_k^{R_k}(t), \dots | t \in s_n) P(t \in s_n)}{P(s_{i_1} \in Neig_k^{R_k}(t), \dots, s_{i_{R_k}} \in Neig_k^{R_k}(t))} \quad (9)$$

Assuming that $P(t \in s_i) = \frac{1}{N} \;\; \forall i$, dropping the denominator term which is common to all speakers and assuming

that the events $\{s_i \in Neig_k^{R_k}(t)\}$ are independent given $\{t \in s_n\}$, $O_c(t)$ and $O_n(t)$ are defined as follows:

$$O_c(t) = \prod_{s_i \in Neig_k^{R_k}(t)} P(s_i \in Neig_k^{R_k}(t) | t \in s_c)$$
$$O_n(t) = \prod_{s_i \in Neig_k^{R_k}(t)} P(s_i \in Neig_k^{R_k}(t) | t \in s_n) \quad (10)$$

For each classifier e_k, the objective function that will be maximized by R_k is defined as follows;

$$\max_{R_k} OF = \sum_{t=1}^{T} f(\delta_{min}(t)) \quad (11)$$

where $f(.)$ is the unit step function and T is the total number of training tokens. We also put the constraint that the correct speaker is not first ranked. This constraint becomes more clear if the algorithm given in Section 5.1 is considered. It can be shown that the constraint enables the satisfaction of item (1) and the term $\delta_{min}(t)$ is used to satisfy the requirement in item (2). The maximum value of objective function, OF_{max} is the number of the training samples for which the correct speaker is in the first rank when the neighbor sets for the optimal R_k value are used.

The optimal thresholds for τ_B and τ_S are experimentally calculated on the cross validation tokens which maximizes the correct classification rate using the combination scheme described in Section 5.1 on the training sessions.

4. THE BODIES OF EVIDENCE AND THEIR BASIC PROBABILITY ASSIGNMENTS

In this section, we will describe how to use the information for the identification of an unseen test token t. During testing, the raw output of each classifier is its ordered neighbor set $Neig_k^{R_k}(t)$ where the first ranked speaker for each e_k will be denoted by s_{k^*}.

Let $\mathbf{W} = \{s_{k^*}\}_{k=1}^{K}$ be defined as the set of the first ranked speakers of all classifiers e_k, $k = 1, 2, \ldots, K$. Let $\overline{Neig}_k^{R_k}(t)$ be defined as follows:

$$\overline{Neig}_k^{R_k}(t) = Neig_k^{R_k}(t) \cap \{\mathbf{W} \cup B_k\}. \quad (12)$$

The intuitive reasoning behind this definition can be stated as follows. Firstly, from the neighbor sets $Neig_k^{R_k}(t)$, the speakers that come out to be first ranked by any one of the classifiers are selected. It should be noted that the speakers in the first ranks are more informative about the correct speaker compared to the speakers placed in second or third or higher ranks. In doing so, the shared evidence by all sources of information is *emphasized*.

As seen from (6), a speaker s_j is in the bad set of the classifier e_k if large number of cross validation tokens of this speaker can *not* be in the top rank of e_k. Similarly, we do not expect the test samples from these speakers to be in the first rank for the given test token. However, we assume that these speakers are among top R_k speakers. Selecting these speakers from the neighbor sets enables the combination scheme to take into account the candidates for correct speaker in these sets.

Let the confusion set of the speaker s_{k^*} be denoted by $\Omega_k^{k^*}$. Then, the confusion set $\Omega_k^{k^*}$ and each speaker in the set $\overline{Neig}_k^{R_k}(t)$ are defined as focal elements and a non zero basic probability value is assigned to each of them as follows.

$$m_k(\Omega_k^{k^*}) = \alpha_k$$
$$m_k(s_n) = \frac{\beta_k O_n(t)}{\sum_{s_j \in \overline{Neig}_k^{R_k}(t)} O_j(t)} \forall s_n \in \overline{Neig}_k^{R_k}(t) \quad (13)$$

where α_k and β_k, $k = 1, 2, \ldots, K$ are design parameters and $O_n(t)$ is defined as follows.

$$O_n(t) = \prod_{s_i \in \overline{Neig}_k^{R_k}(t)} P(s_i \in Neig_k^{R_k}(t) | t \in s_n) \quad (14)$$

Note that the following equality should hold for each classifier.

$$m_k(\Omega_k^{k^*}) + \sum_{s_n \in \overline{Neig}_k^{R_k}(t)} m_k(s_n) = 1 \quad (15)$$

This implies that $\alpha_k + \beta_k = 1$.

The members of the modified neighbor sets, $\overline{Neig}_k^{R_k}(t)$ are naturally used in the selection of focal elements, although they may not include the correct speaker even for the training sessions. However, they are more robust since they are insensitive to the ordering of the speakers in top R_k so tolerant to changes in the ranking of the correct speaker.

On the other hand, $\Omega_k^{k^*}$ always includes the correct speaker for the cross validation tokens. It is assumed that for the test tokens, $\Omega_k^{k^*}$ still contains the correct speaker. Because of this, Θ_k is not defined as a focal element. Only in the cases where there is insufficient training data or noisy *test* samples, there may be a risk that the correct speaker is not in the set $\Omega_k^{k^*}$.

Under the assumption that the training sessions are efficient in generalizing the behavior of the classifiers on the test sessions and considering the fact that the correct speaker may not always exist in the neighbor sets, α_k's should be selected to be larger than β_k's.

Another relationship between α_k and β_k can be obtained from the normalized value of OF. OF, as defined in (11) with optimal R_k value, shows how reliable the neighbors are. So a larger value of OF must give a larger value of β_k.

In the light of these facts and assuming that the correct speaker will always be included in the confusion sets, the relation between α_k's and β_k's becomes

$$\frac{\beta_k}{\alpha_k} = OF'_{max} \quad (16)$$

$OF'_{max} = OF_{max}/P$ where P is the number of tokens which satisfies the constraint in (11). In other words, OF'_{max} is the percentage of the cross validation tokens for which the correct speaker in the top R_k ranks. Using $\alpha_k + \beta_k = 1$ and solving the equations simultaneously, we get

$$\alpha_k = \frac{1}{1 + OF'_{max}} \quad (17)$$
$$\beta_k = 1 - \alpha_k \quad (18)$$

Classifier	Cross Valid. Errors	Test Errors
e_1	326	687
e_2	251	565
Borda count	266	820
Highest rank	289	629
Logistic reg.	246	626
Proposed Method	99	521

Table 1: Average performance of the individual classifiers and different combination methods on SET1-SET4.

5. EVIDENCE COMBINATION AND DECISION MAKING

In this section, the combination scheme is described and the methods that may be used in making the joint decision are presented.

5.1. Hierarchical evidence combination

In order to combine different bodies of evidence, a hierarchical combination scheme is developed. The proposed scheme is efficient in taking into account the strengths and weaknesses of the individual sources of information. Let the speaker s_d denote the joint decision after combination. The scheme is as follows:

Step 1. If $s_{1^*} = s_{2^*} = \ldots = s_{K^*}$, then $s_d = s_{1^*}$. Goto step 5.

Step 2. If $s_{k^*} \in S_k$ for only one classifier e_k, then $s_d = s_{k^*}$. Goto step 5.

Step 3. If there exists at least two classifiers e_k and e_κ such that $s_{k^*} \in S_k$ and $s_{\kappa^*} \in S_\kappa$, then

Step 3.1 If $s_{k^*} = s_{\kappa^*}$, then $s_d = s_{k^*}$. Goto step 5.

Step 3.2 Otherwise select $s_d = s_{k^*}$ where e_k is the classifier with best classification rate on the training sessions. Goto step 5.

Step 4. Using the sources of information and their bpa's as defined is Section 4, apply the Dempster's rule of combination $m = m_1 \oplus m_2 \oplus \ldots \oplus m_K$ and make the joint decision. Goto step 5.

Step 5. End of the algorithm.

The first and second steps of the algorithm are self explanatory. Consider step 3. It may be the case that the first ranked speaker of more than one classifier may be in their sure sets. These speakers may be identical as in the case of step 3.1 in which this speaker is the joint decision. On the other hand, these speakers may be in conflict, that is $s_{k^*} \neq s_{\kappa^*}$ for at least two classifiers. This situation is considered in step 3.2 and s_d is selected as s_{k^*} where e_k is the classifier with best classification rate on the training sessions. When none of the conditions stated above are satisfied, the Dempster's rule of combination is applied. The focal elements of the combined body of evidence are used to decide on the correct speaker. Different criteria that may be used to arrive at a joint decision. In this study, we used the decision rule proposed by Smets *et al* [1].

6. APPLICATION TO THE CLOSED-SET SPEAKER IDENTIFICATION PROBLEM

In order to test the effectiveness of the combination scheme, the proposed method is applied to closed set speaker identification problem. The experiments are conducted on the Polycost database. This database contains around 10 speech sessions recorded by 74 male and 60 female speakers from 14 different countries. From the database, 60 male speakers are grouped into four different sets of 30 speakers and the experiments are done for each set separately. First three sessions of speech files that are used for training are also used for cross validation. The frames of each speech session are partitioned into 20 non overlapping groups. These frame groups are named as *tokens* and each token is treated as a separate training session. Similarly, the test sessions are also blocked into tokens and each token is treated as a separate test session. For each set, there are 1800 cross validation tokens. On the average, there are 3500 test tokens in each speaker set. For the combination problem, two classifiers are developed. For both of the classifiers, 12 Mel frequency cepstral coefficients, i.e. 12-MFCC, and 12 Δ-MFCC coefficients are computed which are concatenated to form a 24 element feature vector per frame [4]. For the first classifier, cepstral mean subtraction (CMS) is applied to the features to minimize the channel variation effects but it is not applied to the second classifier since cepstral means also contain speaker information. Gaussian Mixture modeling is used for characterizing the feature distributions.

7. SIMULATION RESULTS

The performance of the proposed method is tested on four different sets of speakers. These results are given in Table 1 where the errors are averaged over four different sets. The classification results of the proposed algorithm are compared with the well known rank based classifier combination schemes, namely Borda count, highest rank and logistic regression [2]. Experimental results show that the classification system using the proposed sources of information together with the proposed decision combination algorithm surpassed the performance of the individual classifiers and the performance of the other combination methods.

8. REFERENCES

[1] H. Altınçay and M. Demirekler. On the use of supra model information from multiple classifiers for robust speaker identification. *Proceedings of the Eurospeech'99, Budapest, Hungary*, pages 971–974, September 1999.

[2] T. Ho, J. Hull, and S. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:66–75, 1994.

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

[4] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, September 1997.