Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Speech Communication 41 (2003) 531–547

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speaker identification by combining multiple classifiers using Dempster–Shafer theory of evidence

Hakan Altınçay *, Mübeccel Demirekler

Department of Computer Engineering, Eastern Mediterranean University, Gazi Mağusa, KKTC, Mersin 10, Turkey
Department of Electrical and Electronics Engineering, Middle East Technical University, P.K. 06531, Ankara, Turkey

Abstract

This paper presents a multiple classifier approach as an alternative solution to the closed-set text-independent speaker identification problem. The proposed algorithm which is based on Dempster–Shafer theory of evidence computes the first and Rth level ranking statistics. Rth level confusion matrices extracted from these ranking statistics are used to cluster the speakers into model sets where they share set specific properties. Some of these model sets are used to reflect the strengths and weaknesses of the classifiers while some others carry speaker dependent ranking statistics of the corresponding classifier. These information sets from multiple classifiers are combined to arrive at a joint decision. For the combination task, a rule-based algorithm is developed where Dempster's rule of combination is applied in the final step. Experimental results have shown that the proposed method performed much better compared to some other rank-based combination methods.

1. Introduction

Speaker recognition systems using Mel frequency cepstral coefficients for spectral representation of speech signals and Gaussian mixture models (GMM) for speaker modeling have become the state-of-the-art. The telephone channel mismatches and variabilities in the telephone handset characteristics are the main factors having a significant negative effect on the performance of these systems. Robustness against these main factors causing degradations in the performance is one of the major topics attracting the interest of researchers in this field. Different approaches for channel and handset compensation have been proposed (Reynolds, 1997; Heck et al., 2000; Quatieri et al., 1998; Heck and Weintraub, 1997; Quatieri, 2002). Cepstral mean subtraction (CMS) and relative spectral processing (RASTA) (Hermansky and Morgan, 1994; Hermansky et al., 1991) are two most frequently used techniques for channel compensation and HNORM (Reynolds, 1997) is proposed for handset compensation.

In order to improve the robustness against these degradations in speaker identification systems, combination of different spectral representation and speaker modeling approaches have been proposed as an alternative direction where improvements over the individual systems are reported (Furui, 1997; Farrell et al., 1998; Chen et al., 1996;

* Corresponding author. Address: Department of Computer Engineering, Eastern Mediterranean University, Gazi Mağusa, KKTC, Mersin 10, Turkey. Tel.: +90-392-365-0711; fax: +90-392-630-2842.

E-mail addresses: hakan.altincay@emu.edu.tr (H. Altınçay), demirek@eee.metu.edu.tr (M. Demirekler).

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-6393(03)00032-3

Chen and Chi, 1998). In (Altınçay and Demirekler, 1999a), a rule-based and in (Altınçay and Demirekler, 1999b) a Dempster–Shafer theory based combination approach were proposed to combine model sets which were defined by taking into account the classification behavior of the classifiers. Farrell and Mammone combined systems that used vector quantization and neural tree networks for speaker verification using linear and logarithmic opinion pools (Farrell and Mammone, 1995). In (Farrell, 1995), they combined neural tree networks and dynamic time warping based classifiers. Radova and Psutka presented a combination approach for speaker identification where one of the classifiers is based on dynamic time warping and the other is based on using a cepstral distance measure (Radova and Psutka, 1997). They considered voting methods and Borda count method for combination. Brunelli et al. proposed a hybrid rank/measurement level multiple classifier system for person identification using acoustic and visual features (Brunelli and Falavigna, 1995). Chen et al. proposed a novel method based on expectation–maximization in the framework of linear opinion pools and a training approach in designing an associative switch for classifier selection and applied these techniques to speaker identification (Chen et al., 1997). Genoud et al. combined dynamic time warping, second order statistical method and Hidden Markov models based classifiers using weighted majority voting (Genoud et al., 1996). Fredouille et al. described a system where different kinds of information are extracted using several information specific recognizers (Fredouille et al., 2000).

Multiple classifier systems are categorized according to the type of the classifier outputs. The raw classifier outputs may be from the abstract level, class ranking level or measurement level (Ho et al., 1994; Al-Ghoneim and Kumar, 1998; Xu et al., 1992). In the abstract level output case, only the speaker who receives the largest score is the output. In the rank-based classifier combination approach, the output of each classifier is an ordered set of speakers where the ordering is based on their output scores. The measurement level outputs contain the highest amount of information. However, these outputs may have different scales and they may not be directly usable whereas rank-based outputs have the advantage of allowing the combination of arbitrary types of classifiers (Ho et al., 1994).

Ho et al. described three different rank-based combination methods, namely the highest rank, Borda count and logistic regression (Ho et al., 1994). In the highest rank method, each speaker receives a score according to its rank in the classifier output. The score is the largest for the speaker ranked at the top and the speaker ranked at the bottom receives the smallest score. Taking into account the rankings provided by different classifiers, the *combined* score assigned to a speaker is the maximum score it receives from the individual classifiers. The speaker that receives the maximum combined score is selected as the joint decision. In the Borda count combination method, the score assigned to a speaker by a classifier is the number of speakers ranked below the corresponding speaker. The combined score assigned to a speaker 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 (Ho et al., 1994; Al-Ghoneim and Kumar, 1998). The ties in Borda count and highest rank methods which corresponds to the cases where more than one speaker gets the maximal sum score may be broken arbitrarily. For example, the decision maker may randomly select a speaker from those speakers as the final decision. These simple approaches assume that the classifiers have uniform performances.

The raw classifier outputs alone may not be sufficient for achieving improvements since the information about the *strengths* and the *weaknesses* of the individual classifiers should also be used (Ho et al., 1994; Benediktsson and Swain, 1992; Rahman and Fairhurst, 1999). For instance, first ranked speakers of the classifier outputs can be considered as a source of information about the correct speaker since the classifier training is done so as to maximize the correct identification rate. Similarly, the correct speaker is typically placed in the *upper ranked classes* if not placed in the first rank. Hence, knowing that a classifier generally places the correct speaker in the top R ranks for all of the speakers is an important information. On the other hand, knowing that the classifier places the correct speaker in the top P ranks when tested by a particular speaker s_j where $P \ll R$, is an important additional information. Our experiments on the use of multiple classifiers for speaker identification have shown that, when data of a particular speaker is tested, the *set of top R* ranked speakers are not random (Altınçay and Demirekler, 2000). Furthermore, these speakers are in general dependent on the tested speaker. Consequently, *contextual information* (e.g. speaker dependent classifier reliability, global classifier reliability and conflicts among classifiers) about the classifiers should be extracted and quantified using the raw classifier outputs and they should be taken into account during combination (Bloch, 1996). For rank-based combination, Ho et al. proposed the use of logistic regression method which is a rank-based approach taking into account the relative significance of the individual classifiers (Ho et al., 1994). Logistic regression is a modified version of the Borda count method. In this method, 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 this method can be found in (Ho et al., 1994; Pigeon et al., 2000).

In the light of these observations, we propose a novel rank-based classifier combination scheme for a better closed-set text-independent speaker identification system development. In this approach, the speakers are clustered into model sets that carry contextual information about the individual classifiers. These sets were initially introduced in our study described in (Altınçay and Demirekler, 1999a) and they will be named as *information sets* in this context. In this paper, the optimal estimation of these information sets is described and the effectiveness of the use of these sets in the proposed rank-based classifier combination scheme is presented. The proposed algorithm is based on Dempster–Shafer theory of evidence (Shafer, 1976; Voorbraak, 1991; Bhatnagar and Kanal, 1986; Fung and Chong, 1986a; Fung and Chong, 1986b) and it uses the statistical information about these sets in a rule-based manner. The performance of the proposed method is also compared to Borda count, highest rank and logistic regression. In Section 2, the proposed method of rank-based classifier combination and the rule-based decision making algorithm are presented. Sections 3 presents the classifiers and the database used in order to evaluate the proposed combination scheme and gives the experimental results. The conclusions drawn are presented in Section 4 and a brief review of Dempster–Shafer formalism is given in Appendix A.

2. The proposed framework

In this section, the proposed information sets and the combination framework is described. In the first part, the extraction of information from the individual classifiers is described. These information sets are the improved forms of those that were initially published by the authors of this paper in (Altınçay and Demirekler, 1999a). The second part describes a new way of quantifying the extracted information and in the last part, a novel combination algorithm is proposed.

2.1. Information sets extraction

Extraction of information about the behavior of the classifiers is an important part of the multiple classifier systems which corresponds to the estimation of some statistical information about their strengths and weaknesses and the relation between the inputs and their corresponding outputs. In this study, four information sets are defined for the representation of the classifier behavior. These sets are namely, *confusion set*, *bad set*, *sure set* and *neighbor set*.

The confusion set of classifier e_k corresponding to speaker s_j is defined as the set,

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

The term $n_{ij}^{(k,1)}$ is the number of occurrence of s_j in the top rank when classifier e_k tests validation tokens belonging to speaker s_i. Θ is the set of all speakers. During testing an unseen token, if a classifier places s_j in the first rank, the decision maker should take into consideration all speakers in the confusion set of s_j since this speaker appeared at least once in the first rank during testing the validation data of those speakers.

Bad set is the group of speakers for whom only a small percentage of the validation tokens are correctly classified. *The bad set of classifier e_k is defined as the set*,

$$\mathscr{B}_k = \left\{ s_i \left| \frac{n_{ii}^{(k,1)}}{T_i} \leqslant \tau_{\mathrm{B}}, \quad s_i \in \Theta \right. \right\}. \tag{2}$$

T_i denotes the total number of validation tokens that belong to speaker s_i and τ_{B} is a design parameter. If $n_{ii}^{(k,1)}$ comes out to be a small value when all the validation tokens belonging to s_i are considered, it can be argued that the classifier e_k cannot in general correctly classify the corresponding speaker. Hence, the speakers in this set give us information about the *weakness* of the classifier.

The sure set of classifier e_k is defined as the set,

$$\mathscr{S}_k = \left\{ s_i \left| \left(1 - \frac{n_{ii}^{(k,1)}}{\sum_{j=1}^{N} n_{ji}^{(k,1)}} \right) \leqslant \tau_{\mathrm{S}}, \quad s_i \in \Theta \right. \right\}, \tag{3}$$

where N is the total number of speakers. The definition given above indicates that a speaker is in the sure set if the probability of misclassification is very small when the corresponding speaker is first ranked. Hence, the speakers in the sure set give us information about the *strength* of the classifier. In order to ensure this requirement, the threshold τ_{S} should be small.

Neighbor set represents the most likely speakers determined by a classifier e_k for a given speech token t. Determination of the neighbor set of a speech token can be seen in Fig. 1. *Given a speech token t, the Rth level neighbor set,* $\mathrm{Neig}_k^R(t)$, *is defined as the set of speakers that are in the top R ranks when the token is tested by classifier e_k.* The statistics related with the neighbor set is defined as the conditional probability $P(s_j \in \mathrm{Neig}_k^R(t) | t \in s_i)$ which can be approximately calculated as,

$$P(s_j \in \mathrm{Neig}_k^R(t) | t \in s_i) \cong \frac{n_{ij}^{(k,R)}}{T_i}. \tag{4}$$

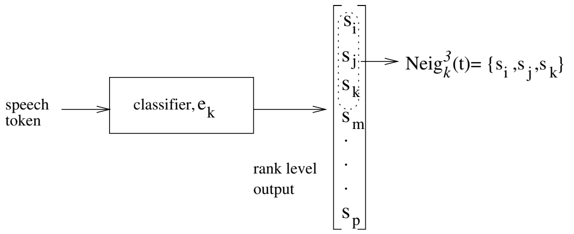

Fig. 1. Determination of the neighbor set corresponding to a speech token for classifier e_k and $R = 3$.

The term $n_{ij}^{(k,R)}$ is the number of occurrence of s_j in the top R ranks when classifier e_k tests validation tokens belonging to speaker s_i. During testing an unseen token, the a posteriori probability that the correct speaker is an arbitrary speaker s_i can be calculated using the total probability theorem as,

$$P(t \in s_i | s_j \in \mathrm{Neig}_k^R(t)) = \frac{P(s_j \in \mathrm{Neig}_k^R(t) | t \in s_i) P(t \in s_i)}{\sum_{l=1}^{N} P(s_j \in \mathrm{Neig}_k^R(t) | t \in s_l) P(t \in s_l)}. \tag{5}$$

Assuming that $T_i = T$ and $P(t \in s_i) = 1/N$, $P(t \in s_i | s_j \in \mathrm{Neig}_k^R(t))$ can be approximately calculated as,

$$P(t \in s_i | s_j \in \mathrm{Neig}_k^R(t)) \cong \frac{n_{ij}^{(k,R)}}{\sum_{l=1}^{N} n_{lj}^{(k,R)}}. \tag{6}$$

2.2. Determination of optimal R, τ_B and τ_S values

From this point on, it is assumed that R is a classifier dependent quantity and it is replaced by R_k for each classifier e_k, $k = 1, \ldots, K$. The criteria that should be taken into account in calculating the optimal values for R_k's are as follows:

1. R_k should be large enough so that the probability that the correct speaker is in the neighbor set is close to 1,
2. R_k should be small enough so that the speakers in the neighbor set are informative about the correct speaker.

Obviously, the best fitting choice of R_k value should be determined over all the validation tokens from the speakers. A reasonable quantity that can be used as a measure of satisfaction of the second requirement is the conditional probability that a token belongs to the correct speaker when the given information is the neighbor set with R_k elements. This probability can be written as $P(t \in s_c | \mathrm{Neig}_k^{R_k}(t))$ where s_c denotes the correct speaker. A larger value for this probability means that we are more confident on s_c when the neighbor set is given. Usually, this probability is small and therefore not reliably computable since the neighbor set alone is not sufficient to determine the correct speaker. A relative value with respect to all other speakers defined as $P(t \in s_c|\mathrm{Neig}_k^{R_k}(t)) - P(t \in s_n|\mathrm{Neig}_k^{R_k}(t))$, where s_n is any other speaker, is the basis for the design of the objective function to be maximized in order to estimate the best fitting R_k value. It is obvious that, for a good classifier, this difference should be large or at least positive. Since the actual values of these probabilities are not reliably computable, we coarsen the objective function as

$$f\left(\min_{\substack{s_n \\ s_n \neq s_c}} \left\{P(t \in s_c|\mathrm{Neig}_k^{R_k}(t)) - P(t \in s_n|\mathrm{Neig}_k^{R_k}(t))\right\}\right), \tag{7}$$

where $f(\cdot)$ is the unit step function. In order to get the best fitting value for all tokens and speakers, this term should be summed over all tokens. After these modifications, the objective function (OF) that must be maximized becomes

$$\mathrm{OF} = \sum_t f\left(\min_{\substack{s_n \\ s_n \neq s_c}} \left\{P(t \in s_c|\mathrm{Neig}_k^{R_k}(t)) - P(t \in s_n|\mathrm{Neig}_k^{R_k}(t))\right\}\right). \tag{8}$$

Let $\mathrm{Neig}_k^{R_k}(t) = \{s_{i_1}, s_{i_2}, \ldots, s_{i_{R_k}}\}$. Using the total probability theorem, we can write

$$\begin{aligned} &P(t \in s_n | s_{i_1} \in \mathrm{Neig}_k^{R_k}(t), \ldots, s_{i_{R_k}} \in \mathrm{Neig}_k^{R_k}(t)) \\ &\quad = \frac{P(s_{i_1} \in \mathrm{Neig}_k^{R_k}(t), \ldots, s_{i_{R_k}} \in \mathrm{Neig}_k^{R_k}(t) | t \in s_n) P(t \in s_n)}{P(s_{i_1} \in \mathrm{Neig}_k^{R_k}(t), \ldots, s_{i_{R_k}} \in \mathrm{Neig}_k^{R_k}(t))}. \end{aligned} \tag{9}$$

Assuming that the events $\{s_i \in \mathrm{Neig}_k^{R_k}(t)\}$ are independent given $\{t \in s_n\}$ and $P(t \in s_i) = 1/N$, $\forall i$, dropping the denominator term which is common to all speakers, $O_c(t)$ and $O_n(t)$ are defined as follows:

$$\begin{aligned} O_c(t) &= \prod_{s_i \in \mathrm{Neig}_k^{R_k}(t)} P(s_i \in \mathrm{Neig}_k^{R_k}(t) | t \in s_c), \\ O_n(t) &= \prod_{s_i \in \mathrm{Neig}_k^{R_k}(t)} P(s_i \in \mathrm{Neig}_k^{R_k}(t) | t \in s_n) \end{aligned} \tag{10}$$

which can be computed using Eq. (4). Now, the objective function can be written as

$$\mathrm{OF} = \sum_t f\left(\min_{\substack{s_n \\ s_n \neq s_c}} \{O_c(t) - O_n(t)\}\right). \tag{11}$$

Experiments on the speaker identification problem have shown that $R_k = 1$ is the optimal value for this function in most of the cases. This is because of the fact that the first speaker is more informative about the correct speaker than the others. However, $R_k = 1$ does not satisfy the first criterion given above that the neighbor sets should include the correct speaker with very high probability. In order to overcome this problem, the objective function is again modified and the tokens that rank the correct speaker as the first are excluded from the sum. Furthermore, if the correct speaker is not in the top R_k ranks, the value of $f(\cdot)$ is considered as zero for the corresponding token even if the argument of the unit step function is greater than zero. These modifications enable the satisfaction of the first criterion.

Note that R_k is the argument of the optimization problem. The value of 'OF' on the other hand is the number of correctly identified validation tokens. Let $\mathrm{OF}_{\max}$ be defined as the value of 'OF' corresponding to the best fitting R_k value. The normalized form of this parameter, $\mathrm{OF}'_{\max}$, is later used in the basic probability assignments on the information sets where,

$$\mathrm{OF}'_{\max} = \frac{\mathrm{OF}_{\max}}{P} \tag{12}$$

and P is the number of tokens used in the summation in Eq. (11).

After the estimation of best fitting R_k values for different classifiers, an $N \times N$ confusion matrix with $n_{ij}^{(k,R_k)}$ values as its ith row and jth column elements is developed for each classifier using the validation tokens as,

$$\mathrm{Conf}_k(R_k) = \begin{bmatrix} n_{11}^{(k,R_k)} & n_{12}^{(k,R_k)} & \dots & n_{1N}^{(k,R_k)} \\ n_{21}^{(k,R_k)} & n_{22}^{(k,R_k)} & \dots & n_{2N}^{(k,R_k)} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N1}^{(k,R_k)} & n_{N2}^{(k,R_k)} & \dots & n_{NN}^{(k,R_k)} \end{bmatrix}. \quad (13)$$

The optimal thresholds for τ_B and τ_S are experimentally determined so that they maximize the correct classification rate of the proposed combination scheme described in Section 2.4 on the validation sessions. The algorithm used can be summarized as follows:

```
/* computation of best fitting $R_k$ value */
for $k = 1,\dots,K$
    Compute $\mathrm{Conf}_k(R_k),\ R_k = 1,2,\dots,N$
    Compute the best fitting $R_k \in \{2,3,\dots,N\}$ using Eq. (11)
end
/* computation of best fitting $\tau_B$ and $\tau_S$ values */
for $\tau_B = 0$ to 1 step (1/60)
    for $\tau_S = 0$ to 0.5 step 0.01
        Run the combination algorithm given in Section 2.4 on the
        validation tokens and record the recognition rate for each $\tau_B$ and $\tau_S$
    end
end
/* computation of information sets using best fitting $\tau_B$ and $\tau_S$ values */
for $k = 1,\dots,K$
    Compute $\mathrm{Conf}_k(R_k)$, $\mathscr{B}_k$, $\mathscr{S}_k$ and $\Omega_k^j$, $j = 1,\dots,N$
end
```

In addition to these, the individual classification rate of each classifier on the validation tokens denoted by perf_k is computed for being used during testing. A summary of the information extracted from the individual classifiers is given in Table 1. The next two subsections are devoted to the description of testing an unseen speech token.

2.3. Testing an unseen token: basic probability assignments on the information sets

When testing an unseen token, the determination of the focal elements starts with a modification of the neighbor sets provided by the classifiers. Let s_{k^*} denote the first ranked speaker obtained during testing by classifier e_k and $\boldsymbol{W} = \{s_{k^*}\}_{k=1}^{K}$. Then, the modified neighbor set is obtained as,

$$\overline{\mathrm{Neig}}_k^{R_k}(t) = \{\mathrm{Neig}_k^{R_k}(t) \cap \boldsymbol{W}\} \cup \{\mathrm{Neig}_k^{R_k}(t) \cap \mathscr{B}_k\}. \quad (14)$$

The intuitive reasoning behind this modification is to emphasize the evidence shared by all sources of information contained in $\boldsymbol{W} = \{s_{k^*}\}_{k=1}^{K}$ and similarly the speakers in $\mathscr{B}_k$ since bad set speakers seldom appear in the first rank.

Another set that may contain the correct speaker is the confusion set of the first ranked speaker denoted by $\Omega_k^{k^*}$. However, the first ranked speakers given by each classifier are sensitive to mismatch conditions and this directly effects the confusion sets. On the other hand, neighbor sets are more robust to mismatch conditions since the ordering

Table 1
A summary of the information extracted from individual classifiers

Extracted information	Description
Confusion sets, Ω_k^j	Gives the set of speakers for which s_j may come out to be the most likely speaker. It is speaker and classifier dependent
Bad set $\mathscr{B}_k$	Gives the set of speakers that are not placed in the first rank by classifier e_k. A bad set is determined for each classifier
Sure set $\mathscr{S}_k$	Gives the set of speakers that are almost always placed in the first rank by classifier e_k. A sure set is determined for each classifier
R_k	Optimal cardinality of the neighbor set of classifier e_k
R_kth level confusion matrix, $\mathrm{Conf}_k(R_k)$	An $N \times N$ matrix containing $n_{ij}^{(k,R_k)}$ as its ith row and jth column elements
perf_k	Correct identification performance of the classifier e_k on the validation tokens

of the speakers in these sets is not considered. It should be noted that, in this context, the mismatch conditions corresponds to the cases where the first ranked classifier outputs for a given test utterance are different from those obtained for the same speaker during validation. As an example, if a speaker whose model was placed in the first rank for all validation tokens comes out as the second ranked speaker for some test tokens, it is captured in the neighbor sets although it may be missed in the confusion sets. Hence, in order to avoid the estimation of unreliable statistics, $\Omega_k^{k^*}$'s are directly used as focal elements where the neighbor sets that are more robust are further refined to sets that contain single elements, $\{s_i\}$. The major gain in this refinement is the elimination of the ties after combination. The combination of neighbor sets directly was observed to provide some bigger speaker sets providing the same combined bpa for several speakers leading to a more critical decision making process.

Since $\mathrm{Bel}(\Theta) = 1$, we have,

$$m_k(\Omega_k^{k^*}) + \sum_{s_\mathrm{n} \in \overline{\mathrm{Neig}_k^{R_k}}(t)} m_k(\{s_\mathrm{n}\}) = 1. \tag{15}$$

To compute the basic probability assignments, let us define,

$$\begin{aligned} m_k(\Omega_k^{k^*}) &= \alpha_k, \\ \sum_{s_\mathrm{n} \in \overline{\mathrm{Neig}_k^{R_k}}(t)} m_k(\{s_\mathrm{n}\}) &= \beta_k, \end{aligned} \tag{16}$$

where α_k and β_k, $k = 1, 2, \ldots, K$ are design parameters such that $\alpha_k + \beta_k = 1$. Let $\overline{\mathrm{Neig}_k^{R_k}}(t) = \{s_{i_1}, s_{i_2}, \ldots, s_{i_p}\}$, where $p \leqslant R_k$ is the cardinality of $\overline{\mathrm{Neig}_k^{R_k}}(t)$. Once α_k and β_k are computed, it remains to distribute the belief assigned to $\overline{\mathrm{Neig}_k^{R_k}}(t)$, β_k, over the speakers in the set. Intuitively, each $m_k(\{s_\mathrm{n}\})$ term in Eq. (16) should be proportional with $P(t \in s_\mathrm{n} | s_{i_1} \in \mathrm{Neig}_k^{R_k}(t), \ldots, s_{i_p} \in \mathrm{Neig}_k^{R_k}(t))$. Assuming the conditional independence among the events $\{s_i \in \overline{\mathrm{Neig}_k^{R_k}}(t)\}$ given $\{t \in s_\mathrm{n}\}$ and considering Eq. (9), $m_k(\{s_\mathrm{n}\})$ becomes proportional with the a priori probabilities as,

$$m_k(\{s_\mathrm{n}\}) \propto \prod_{s_i \in \overline{\mathrm{Neig}_k^{R_k}}(t)} P(s_i \in \mathrm{Neig}_k^{R_k}(t) | t \in s_\mathrm{n}). \tag{17}$$

Defining $\overline{O}_\mathrm{n}(t)$ as,

$$\overline{O}_\mathrm{n}(t) = \prod_{s_i \in \overline{\mathrm{Neig}_k^{R_k}}(t)} P(s_i \in \mathrm{Neig}_k^{R_k}(t) | t \in s_\mathrm{n}) \tag{18}$$

$m_k(\{s_\mathrm{n}\})$ becomes,

$$m_k(\{s_\mathrm{n}\}) = \frac{\overline{O}_\mathrm{n}(t)}{\overline{O}_\mathrm{tot}} \beta_k, \tag{19}$$

where

$$\overline{O}_\mathrm{tot} = \sum_{s_\mathrm{n} \in \overline{\mathrm{Neig}_k^{R_k}}(t)} \overline{O}_\mathrm{n}(t). \tag{20}$$

The computation of α_k and β_k values means the computation of their relative values since their sum is 1. Another relationship between α_k and β_k is obtained from the normalized value of the objective function defined in Eq. (11) with optimal R_k value. Since 'OF' gives the number of correctly identified tokens using neighbor sets, a larger values of 'OF' must give a larger β_k. In the light of these facts and assuming that the correct speaker is always included in the confusion sets, the relation between α_k's and β_k's becomes,

$$\alpha_k + \beta_k = 1, \tag{21}$$

$$\frac{\beta_k}{\alpha_k} = \mathrm{OF}'_\mathrm{max}, \tag{22}$$

where $\mathrm{OF}'_\mathrm{max}$ is the percentage of correctly classified tokens as defined in Eq. (12). Solving the above equations simultaneously, we get

$$\alpha_k = \frac{1}{1 + \mathrm{OF}'_\mathrm{max}}, \tag{23}$$

$$\beta_k = 1 - \alpha_k. \tag{24}$$

Section 2.4 describes the final step of the proposed approach where the information sources together with their basic probability assignments are used to make a joint decision.

2.4. Testing an unseen token: information combination and decision making

Fig. 2 gives the overall block diagram of the proposed algorithm for two classifiers case.

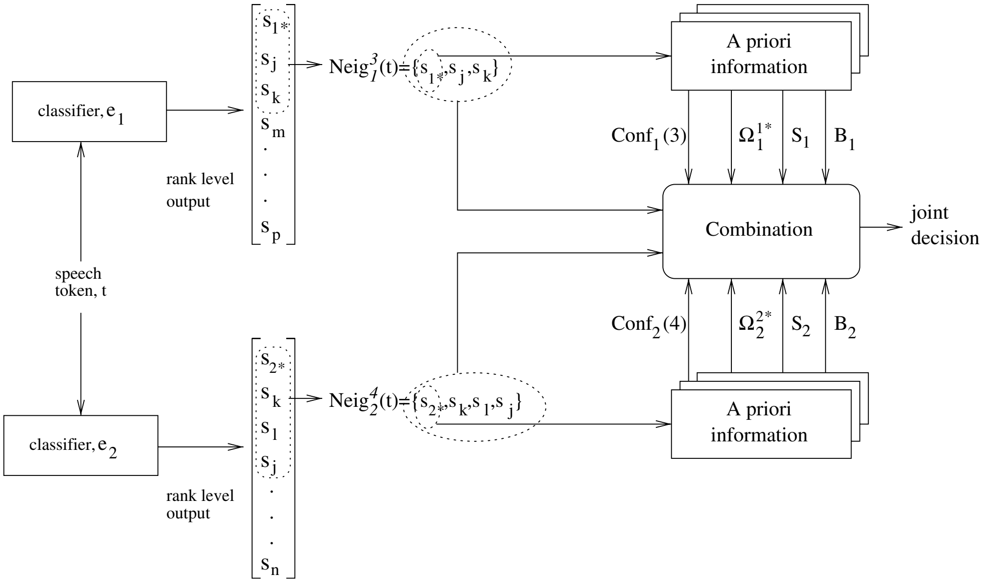

Fig. 2. Overall block diagram of the proposed multiple classifier speaker recognition system when *testing* an unknown token *t* for two classifiers case. It is assumed that $R_1 = 3$ and $R_2 = 4$.

The combination algorithm is given below where the final decision is denoted by s_{d}.

Step 1. If $s_{1^*} = s_{2^*} = \cdots = s_{K^*}$, then $s_{\mathrm{d}} = s_{1^*}$. Goto step 5.
Step 2. If $s_{k^*} \in \mathscr{S}_k$ for only one classifier e_k, then $s_{\mathrm{d}} = s_{k^*}$. Goto step 5.
Step 3. If there are at least two classifiers e_k and e_m such that $s_{k^*} \in \mathscr{S}_k$ and $s_{m^*} \in \mathscr{S}_m$, then
  Step 3.1 If $s_{k^*} = s_{m^*}$, then $s_{\mathrm{d}} = s_{k^*}$. Goto step 5.
  Step 3.2 If $\mathrm{perf}_k > \mathrm{perf}_m$, select $s_{\mathrm{d}} = s_{k^*}$, otherwise $s_{\mathrm{d}} = s_{m^*}$. Goto step 5.
Step 4. Using the sources of information and their basic probability asignments as defined is Section 2.3, apply Dempster's rule of combination $m = m_1 \oplus m_2 \oplus \cdots \oplus m_K$. Goto step 5.
Step 5. End of the algorithm.

In step 1, it is checked whether the classifiers reach at a consensus in their first ranked speakers. In this case, this speaker is selected as the joint decision. Consensus among the information sources is generally accepted as an indication of correct decision from all classifiers (Battiti and Colla, 1994).

Step 2 corresponds to the case where the first ranked speaker of only one of the classifiers is an element of its sure set. This step is used in the case of a conflict among the classifiers and the combination algorithm uses the strength of the sure sets. Since the probability of misclassification is very small when a speaker in the sure set is first ranked, the speaker is selected as the joint decision.

There may be cases where the first ranked speakers of several classifier are in their sure sets. If these speakers are the same, then it is selected as the joint decision as indicated in step 3.1. On the other hand, first ranked sure set speakers may be in conflict, that is $s_{k^*} \neq s_{m^*}$ for at least two classifiers. This situation is considered in step 3.2. Among the classifiers having their most likely speakers in their sure sets, s_{d} is selected as the first ranked speaker of the classifier with the highest classification rate on the validation tokens.

When none of the conditions stated above are satisfied, Dempster's rule of combination is applied. The focal elements of the combined body of evidence and their basic probability assignments are used to decide on the correct speaker. In this study, Decision rule 3 given in Appendix A is used. As a final remark, notice that the Dempster–

Shafer formalism is applied only in the cases where the classifiers are in conflict and the top ranked speaker is not in the sure set of at least one classifier.

3. Experiments

In order to evaluate the proposed classifier combination scheme, some experiments are conducted on six different speaker sets. In the sequel, the classifiers and the database used in the simulation experiments on the proposed method are described and the experimental results are presented.

3.1. The architecture and training of the classifiers

The experiments are conducted on the POLYCOST database (Melin and Lindeberg, 1997). On the average, this database contains 10 sessions of speech records for each 74 male and 60 female speakers. Each session contains 14 items; four repetitions of a seven-digit client code, five ten-digit sequences, two fixed sentences, one international phone number and two items with speech in the subject's mother tongue. For each speaker, the recording sessions were spread over the planned period between February and April 1996 with a minimum spacing of three days between the sessions. The speech recording is done on international telephone lines. The speakers used their phone sets in their countries and, on an average, 80% of the speakers called from the same phone set in all sessions. In our experiments, only the mother tongue utterances are used. A free-text mother tongue utterance from each of the first three sessions are used for training and the two mother tongue utterances in all of the sessions starting from session five are used for testing. The experiments are conducted on four arbitrarily selected sets of male speakers SET1,…, SET4 and a set of female speakers, SET5. Furthermore, a set containing both female and male speakers, SET6, is used to investigate the relation between the performance improvement and token length. The speakers and the total number of test tokens in each of these sets are given in Table 3.

Speech utterances are blocked into frames of length 20 ms with 10 ms overlapping for the short-time spectral analysis and automatically segmented into four broad sound classes as voiced, unvoiced, transition and silence. The segmentation is based on the energy and zero crossing counts of the speech frames. Some thresholds are experimentally determined for these measures and they are used for frame by frame segmentation of the speech records. The transition frames are determined by examining the norm of the difference of the 12 Mel frequency cepstral coefficients (12-MFCC) corresponding to the preceding and following frames.

The segmentation is followed by speaker model training. For each frame, 12-MFCC that were computed during the segmentation are concatenated with their deltas to form 24 element feature vectors per frame (Campbell, 1997). These feature vectors are used to train three GMM for each speaker as λ^i_{voi}, λ^i_{unv} and $\lambda^i_{\mathrm{trans}}$ using voiced, unvoiced and transition regions separately (Reynolds and Rose, 1995). The silence regions in the training records of the speakers are used to train a single GMM, λ_{sil}, which is common to all speakers. During testing, the likelihood of the speaker model $\lambda^i = \{\lambda^i_{\mathrm{voi}}, \lambda^i_{\mathrm{unv}}, \lambda^i_{\mathrm{trans}}, \lambda_{\mathrm{sil}}\}$ is computed as,

$$P(\bar{x}|\lambda^i) = \max\{P(\bar{x}|\lambda^i_{\mathrm{voi}}), P(\bar{x}|\lambda^i_{\mathrm{unv}}), P(\bar{x}|\lambda^i_{\mathrm{trans}}), P(\bar{x}|\lambda_{\mathrm{sil}})\}. \tag{25}$$

For speaker recognition applications, it was observed that the best choice of the number of mixtures depends on the length of the training data (Reynolds and Rose, 1995) and diagonal covariance matrix GMMs performed better than those using full covariance matrices (Reynolds et al., 2000). In our experiments, owing to limited amount of training data (approximately 20 s in each utterance) 16 mixtures and diagonal covariance matrices are used for each GMM. The mixtures are initialized using random mean selection followed by a single iteration *k*-means clustering algorithm and the GMM parameters are then optimized using the expectation–maximization algorithm.

Table 2
The classifiers used in the simulation experiments

Classifier	Feature set	CMS	Speaker model
e_1	12-MFCC + Δ-MFCC	√	GMM
e_2	12-MFCC + Δ-MFCC		GMM

Note that, CMS is not applied for the second classifier.

Table 3
The speakers included in different speaker sets and the total number of test tokens in these sets

Test sets	Speakers included	Total test tokens
SET1	m001,...,m030	3460
SET2	m031,...,m060	3560
SET3	m015,...,m029,m031,...,m045	3420
SET4	m001,...,m015,m046,...,m060	3520
SET5	f001,...,f030	3380
SET6	f001,...,f015, m001,...,m015	–

In SET6, the number of test tokens is not fixed.

The proposed system involves two classifiers. The basic properties of the classifiers are given in Table 2. Notice that the only difference between e_1 and e_2 is CMS. CMS is generally applied to the features in order to minimize the channel variation effects but cepstral means also contain speaker information (Gish and Schmidt, 1994; Atal, 1974; Soong and Rosenberg, 1988). Because of this, it is preferred to use them separately in different classifiers.

In order to train reliable confusion matrices, a sufficient number of validation utterances should be used and each session of the validation data should contain a wide range of acoustical sound classes. In order to achieve this, the speech utterances used for training are also used for validation as follows: The frames of each 20 s training utterance are partitioned into 20 non-overlapping groups. These frame groups are named as *tokens*. Three mother tongue utterances, one from each of the first three sessions of the database, are used during training and validation and this corresponds to a total of 60 tokens for each speaker. For a 20 s training utterance, each token has a fixed length of 1 s and contains 100 speech frames. Each time leaving out the tokens from one utterance for the validation (i.e. 20 tokens) and using the tokens from the remaining two utterances for model training (i.e. 40 tokens), all 60 tokens from three utterances are used for validation. This kind of approach is known as cross-validation (Gish and Schmidt, 1994; Xu et al., 1992). Since each set contains 30 speakers, there are totally $30 \times 60 = 1800$ tokens used for training and validation. During testing, the mother tongue utterances in all of the test sessions starting from session five are used and they are similarly partitioned into 20 tokens.

3.2. Results

The proposed method is evaluated on six different sets of speakers. For brevity, the estimated values for the design parameters are given only for SET1 in Table 4. Some comments ought to be made on the simulation results. Notice that $\tau_S = 0$. This means that the first ranked speaker of a classifier is surely decided as the correct speaker only if there is no risk of misclassification which is also intuitively reasonable. Notice from Table 4 that a speaker may be in both the sure set and the bad set of a classifier. This may seem to be a contradictory statement but actually it is not. A speaker s_i is in the sure set only if there is no risk in

Table 4
The estimated values of the design parameters on SET1

Parameter	Estimated parameter value
R_1	6
R_2	12
α_1	0.59
α_2	0.59
β_1	0.41
β_2	0.41
τ_S	0.0
τ_B	0.5
$\mathscr{B}_1$	$\{s_4, s_{11}\}$
$\mathscr{B}_2$	$\{s_1, s_4, s_{11}, s_{25}\}$
$\mathscr{S}_1$	$\{s_4, s_6, s_{30}\}$
$\mathscr{S}_2$	$\{s_1, s_2, s_4, s_6, s_7, s_{10}, s_{11}, s_{21}, s_{25}\}$

deciding on that speaker when it is first ranked. This is equivalent to saying $n_{ii}^{(k,1)}$ is much greater than the other elements of the ith *column* of the first level confusion matrix. On the other hand, the bad set takes into account the *misclassification* behavior of the classifier. This is observed when $n_{ii}^{(k,1)}$ is not large enough when compared to the other elements of the ith *row* of the first level confusion matrix. So, for SET1, $s_4 \in \mathscr{S}_1$ and $s_4 \in \mathscr{B}_1$ means that the classifier was *ineffective* in correctly classifying the validation tokens from s_4 but when the first ranked speaker comes out to be s_4, there is no risk in deciding on this speaker. Fig. 3 illustrates the classification errors, $(1 - \mathrm{OF}/P)$, when the correct speaker is not ranked the first and the most likely R speakers are used to decide on the correct speaker (refer to Section 2.2, Eq. (11)), as a function of different R values respectively for SET1. Remember that P is the number of validation tokens for which the correct speaker is not ranked the first by the classifier.

Table 5 summarizes the testing errors on four different sets of speakers for the individual classifiers, three other combination methods and the proposed approach. In the table, it can be seen that the improvement is much larger on SET1. For this speaker set, the cardinality of the bad set is much smaller compared to the bad sets of other speaker sets and the improvement is nearly indirectly proportional to these cardinalities for all four speaker sets. Actually, such a relation between the cardinality of the bad sets and the improvement in the classification accuracy is reasonable. A speaker lies in the bad set if its performance is very poor on the validation tokens.

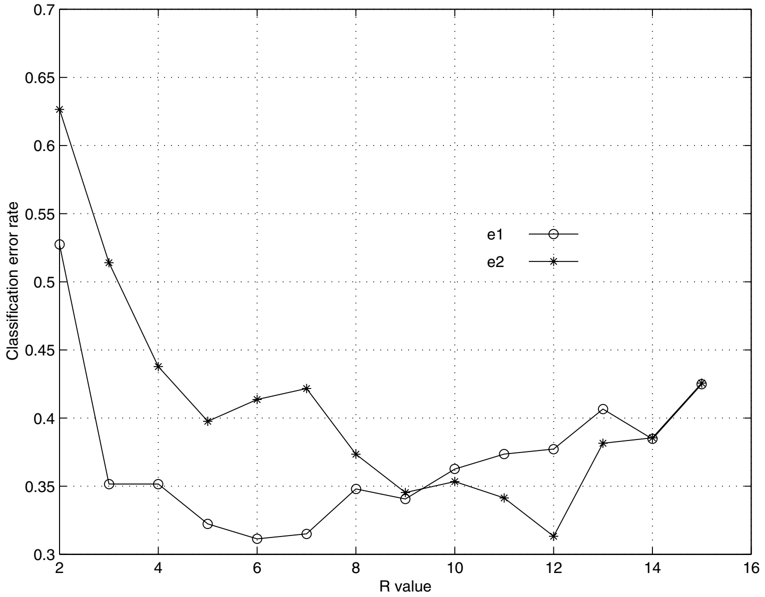

Fig. 3. The percentage of classification errors for the cross validation tokens when the neighbor sets are used and the correct class is not the first ranked in the raw data. The results are for SET1.

Table 5
Identification *errors* of the individual classifiers, three different combination methods and the proposed method on four sets of speakers

Classifier	SET1	SET2	SET3	SET4	Average
e_1	517	870	542	820	687
e_2	397	783	465	614	565
Borda count	653	1124	697	806	820
Highest rank	457	837	504	718	629
Logistic regression	479	810	498	718	626
Proposed method	**299**	**758**	**433**	**595**	**521**
Improvement (%)	25	3	7	3	8

The last row gives the percent reduction in the errors. The rightmost column gives the average number of errors over four different speaker sets and the average improvement in the testing errors over the best individual classifier.

In such a case, the speaker models should be suspected as not representing the speaker characteristics and behaving in a rather random manner. On the average of the first four speaker sets, there is 8% reduction in the testing errors compared to the best individual classifier. When the system is tested by the validation tokens that were used for training, 60% reduction in the errors is achieved on the average. The reduction in the errors of the test data came out to be 7% for SET5 involving only the female speakers which is comparable to that of male speakers.

The effect of the token length on the percent reduction of the errors is studied on SET6. The *number* of tokens is selected as 12, 14, 16, 18 and 20. The token length is maximum in the case of 12 tokens and it decreases as the speech utterances are divided into higher number of tokens. The precent reduction in the errors is given in Fig. 4. As seen in the figure, the improvement achieved is larger for longer tokens.

Different steps of the algorithm are analyzed for their performance. On the average over the first four sets of speakers, the first step of the algorithm

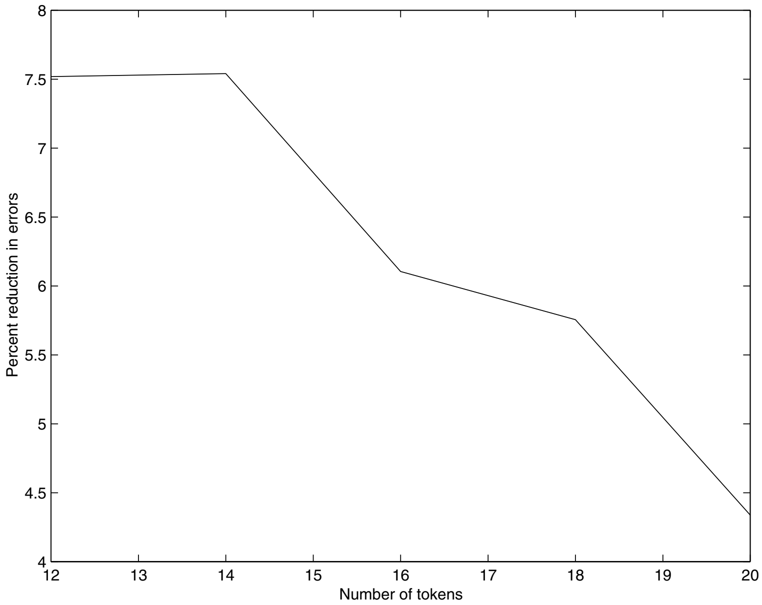

Fig. 4. The relation between the percent improvement in the classification errors and the number of tokens for test utterances. It should be noted that as the number of tokens used increases, their lengths decreases.

is used to make the final decision for 78% of the tokens. This means that, in more than 20% of cases, the classifiers were in conflict. This is an interesting result because, although the only difference between the classifiers is in their feature vectors, their classification behavior may change to such an extent. Ninety six percentage of the decisions made in this step are correct. The final decision is made for 2% of the tokens in steps two and three where sure sets are used and 35% of these tokens are correctly classified. This step corresponds to the case where there is a conflict in the first ranked speakers of the classifiers but one or more of the classifiers are sure about their most likely speakers. In step four, decisions are made for 20% of the tokens and 50% of these decisions are correct. Considering also the cases where the most likely speakers from both of the classifiers are incorrect, the neighbor sets provide some robustness against the variabilities between the behavior of the classifiers on the test tokens and the learned behavior on the validation tokens.

Although its performance was promising for the cross validation tokens, the logistic regression method was not able to perform better compared to the best individual classifier for the case of test tokens. This is not surprising since Ho et al. (1994) reported that improvement in the top choice is usually achieved *only* if three or more classifiers are used in the combination task. It was suggested in (Ho et al., 1994) that, for the case of two classifiers, correlation in errors prevents improvements in the top rank choices. The correlation in errors, which corresponds to the cases where both of the individual classifiers make errors, can only be avoided if they are learned during training. In the proposed algorithm, some effort is spent on this problem. For example, correlated errors generally occur for the speakers in the bad sets of both of the classifiers. For those speakers, both of the classifiers are expected to make errors and such errors cannot be avoided in the combination methods like the highest rank. In the proposed approach, the algorithm carefully deals with such cases by considering the bad speakers in top R_k speakers in defining the modified neighbor sets as seen in Eq. (14). It may be concluded that the proposed combination scheme in this paper is more efficient in avoiding the correlated errors between the classifiers by putting some emphasis on the speakers which are not expected to be correctly classified.

As a final remark, it should be stated that the computational complexity of the proposed system is proportional to the number of classifiers used. As a matter of fact, the combination time is nearly negligible compared to the GMM likelihood computation in the individual classifiers. In addition, the preprocessing stage of the second classifier only has an additional CMS operation. Thus, the testing time for the proposed multiple classifier system with two 16 component GMM classifiers is nearly equal to the testing time of a single 32 component GMM classifier.

4. Conclusions

A method of classifier combination using rank level classifier outputs has been developed for speaker identification. In the proposed method, contextual information about the classifiers is represented in terms of some information sets. These sets are optimally estimated and then they are used in a rule-based combination algorithm. In the last step of this decision making algorithm, Dempster–Shafer evidence theory based formalism is used which is highly suitable for processing uncertain information. In the described method, the strengths and weaknesses of the individual classifiers are learned and these facts are used during combination. The use of the set of top R speakers was also proposed as a source of information about the correct speaker. The experiments on speaker identification have shown that the proposed combination scheme effectively learned these ranking statistics about the individual classifiers and these statistics were effective in the combination process.

The way of extracting contextual information about the classifiers is shown to be an important issue in the classifier combination task. Actually, a combination approach is expected to be an efficient one only if the information extracted about the individual classifiers takes into account the nature of the task under concern.

The analysis of the proposed algorithm revealed some important facts about the issues to be further studied in speaker identification by using multiple classifier systems. Although the only difference between the classifiers is in their feature sets, the classification behavior changed a lot, providing the same most likely speaker only in approximately 78% of the cases. Therefore, the classification behavior of speaker identification systems is highly sensitive to the preprocessing and modeling techniques. Improving the robustness against channel or handset variablities may result in increased matching between the validation and test behaviors of the classifiers. Such efforts may lead to better multiple classifier speaker identification systems.

For the speaker identification task, plenty of different features and classification algorithms are proposed in the literature. It is still very difficult to implement a unique classifier which provides sufficient performance for practical applications. The multiple classifier approach is a candidate approach for this purpose. In this study, we preferred to combine a channel noise robust classifier where CMS is applied and a classifier where mean subtraction is not applied. As a further research, a classifier which is effective in compensating the effects of using handsets with different characteristics can also be considered.

We believe that the use of multiple information sources has the potential of developing better speaker identification systems. As stated in recent studies, information fusion can be considered as a future direction in speaker recognition applications (Reynolds et al., 2000; Doddington et al., 2000).

Appendix A. Dempster–Shafer formalism: a brief review

Let Θ denote the finite set of all speakers in the given closed-set speaker identification problem which is also known as the *frame of discernment* (Shafer, 1976; Bhattacharya, 2000; Fung and Chong, 1986a). Assume that A and B denote two arbitrary speaker sets where $A \subset \Theta$ and $B \subset \Theta$. *Basic probability assignment* (*bpa*) denoted by $m(\cdot)$ is defined as a function from all the subsets of Θ to the unit interval [0,1]. $m(\phi) = 0$, meaning that the basic probability assigned to the empty set is zero and $\sum_{A \subseteq \Theta} m(A) = 1$. $m(A)$ is a measure of the support *exactly* assigned to A and it is ignorant about the precise division of support among the subsets of the set A. The definition of this function is specific to the application considered.

In order to calculate the total *belief* assigned to a set A, the basic probabilities assigned to all subsets of A are added. The *belief function* $\mathrm{Bel}(\cdot)$ is defined as a function that maps any subset A of Θ into the interval [0,1] such that $\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B)$, $\mathrm{Bel}(\phi) = 0$ and $\mathrm{Bel}(\Theta) = 1$. The subsets of Θ that have non-zero basic probability values are called *focal elements* and the union of all the focal elements is called the *core* of a belief function.

The classifier outputs are then combined by using the Dempster's rule (Shafer, 1976; Shafer and Logan, 1987). Consider two sets of speakers A and B with basic probabilities $m_1(A)$ and $m_2(B)$. The basic probability of their conjunction $C = A \cap B$ is proportional to $m_1(A) \times m_2(B)$. There may be several pairs of sets A and B whose conjunction is equal to a given set C. The total basic probability assigned to the set C will be the sum of all such contributions. However, if $C = A \cap B$ is an empty set, the combination of the corresponding *bpa*'s will support the empty set ϕ. This difficulty can be resolved by multiplying the *bpa* of the non-empty sets with a normalization factor and this corresponds to distributing the basic probability assigned to empty sets among the non-empty sets. The formulation of the combination process is as follows:

$$(m_1 \oplus m_2)(C) = \frac{\sum_{A \cap B = C} m_1(A) m_2(B)}{1 - \sum_{A \cap B = \phi} m_1(A) m_2(B)}. \qquad \text{(A.1)}$$

The numerator of Eq. (A.1) represents the sum over all conjunctions of arguments which support C. The denominator is the normalization coefficient obtained from the mass assigned to the contradictory arguments. Note that the combination is not defined for the case where the cores of the two *bpa*'s are disjoint, that is, when there is total conflict between the two information sources. In this case, the denominator of Eq. (A.1) becomes zero leading to an undefined result of combination.

The focal elements $\{F_j\}_{j=1}^{J}$ of the combined body of evidence $m_{\mathrm{c}} = m_1 \oplus m_2 \oplus \cdots \oplus m_K$ are used in making the joint decision on the correct speaker. There are a number of decision rules that are in general used for this purpose (Xu et al., 1992; Shafer, 1976; Smets and Kennes, 1994; Hegarat-Mascle et al., 1998). The rules can be summarized as follows:

Decision rule 1. The joint decision is the speaker s_{d} which gets the maximum belief value where,

$$s_{\mathrm{d}} = \arg\max_{s_i} \mathrm{Bel}(s_i) = \arg\max_{s_i} \sum_{F_j \subseteq \{s_i\}} m(F_j)$$
$$= \arg\max_{s_i} m(\{s_i\}).$$

The belief function represents the minimum uncertainty about the speaker under concern.

Decision rule 2. The joint decision is the speaker s_{d} where

$$s_{\mathrm{d}} = \arg\max_{s_i} \sum_{F_j \cap \{s_i\} \neq \phi} m(F_j) = \arg\max_{s_i} \sum_{\{s_i\} \subseteq F_j} m(F_j).$$

This decision rule is also known as *Plausibility*, $Pl(\cdot)$ where $Pl(A)$ represents the total evidence that is contained in the sets overlapping with A (Voorbraak, 1991). It also represents the maximum uncertainty about the speaker under concern.

Decision rule 3. The joint decision is the speaker s_{d} with maximum pignistic probability as defined by Smets and Kennes (1994):

$$s_{\mathrm{d}} = \arg\max_{s_i} \sum_{\{s_i\} \subseteq F_j} \frac{m(F_j)}{\mid F_j \mid}$$

which reflects the sum of the scaled evidences of the sets containing the set under concern. The scaling is done by the cardinalities of the sets.

An Example. Consider a four speaker closed-set speaker identification problem where $\Theta = \{s_1, s_2, s_3, s_4\}$ and two information sources with the corresponding basic probability assignments $m_1(\cdot)$ and $m_2(\cdot)$ respectively. Assume that these functions are defined on the focal elements of the individual information sources as,

$$m_1(\{s_1, s_2\}) = 0.5, \qquad m_2(\{s_1, s_3\}) = 0.6,$$
$$m_1(\{s_1, s_2, s_3, s_4\}) = 0.5, \qquad m_2(\{s_1, s_2, s_3, s_4\}) = 0.4.$$

Note that, $\{s_1, s_2\}$ and $\{s_1, s_2, s_3, s_4\}$ are the focal elements of the first information source. Using the definition,

$$\begin{aligned}\mathrm{Bel}_1(\{s_1, s_2, s_3\}) &= \sum_{B \subseteq \{s_1, s_2, s_3\}} m_1(B) = m_1(\{s_1, s_2\}) = 0.5,\\ \mathrm{Bel}_2(\{s_1, s_2, s_3\}) &= \sum_{B \subseteq \{s_1, s_2, s_3\}} m_2(B) = m_2(\{s_1, s_3\}) = 0.6.\end{aligned} \tag{A.2}$$

Using Dempster's rule of combination, the resultant basic probability assignments $m_{\mathrm{c}}(\cdot) = m_1 \oplus m_2$ reflecting the basic probability values of the combined body has the following form:

$$\begin{aligned} m_{\mathrm{c}}(\{s_1\}) &= 0.3,\\ m_{\mathrm{c}}(\{s_1, s_2\}) &= 0.2,\\ m_{\mathrm{c}}(\{s_1, s_3\}) &= 0.3,\\ m_{\mathrm{c}}(\{s_1, s_2, s_3, s_4\}) &= 0.2.\end{aligned} \tag{A.3}$$

The sets $F_1 = \{s_1\}$, $F_2 = \{s_1, s_2\}$, $F_3 = \{s_1, s_3\}$ and $F_4 = \{s_1, s_2, s_3, s_4\}$ are the focal elements that result from the conjunction of the focal elements of the individual information sources using Eq. (A.1). Similarly, the belief values of *any set of speakers* can be calculated using the combined basic probability assignments. For instance, $\mathrm{Bel}_{\mathrm{c}}(\{s_1, s_2\}) = m_{\mathrm{c}}(\{s_1\}) + m_{\mathrm{c}}(\{s_1, s_2\}) = 0.5$. Using the rule proposed by Smets and Kennes, the joint decision will be speaker s_1 since it maximizes the value of the corresponding objective function, taking the value $\sum_{\{s_i\} \subseteq F_j} (m_{\mathrm{c}}(F_j)/|F_j|) = 0.3 + 0.2/2 + 0.3/2 + 0.2/4 = 0.6$.

References

Al-Ghoneim, K., Kumar, B.V.K.V., 1998. Unified decision combination framework. Pattern Recognition 31 (12), 2077–2089.

Altınçay, H., Demirekler, M., 1999a. Use of model confusion information for speaker identification: A rule based approach. In: Proceedings of the IEEE-EURASIP Nonlinear

Signal and Image Processing (NSIP'99), Antalya, Turkey, June 1999, pp. 321–325.

Altınçay, H., Demirekler, M., 1999b. On the use of supra model information from multiple classifiers for robust speaker identification. In: EUROSPEECH Proceedings, September, pp. 971–974.

Altınçay, H., Demirekler, M., 2000. A novel rank-based classifier combination scheme for speaker identification. In: IEEE-ICASSP Proceedings, June.

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55 (6), 1304–1312.

Battiti, R., Colla, A.M., 1994. Democracy in neural nets: Voting schemes for classification. Neural Networks 7 (4), 691–707.

Benediktsson, J.A., Swain, P.H., 1992. Consensus theoretic classification methods. IEEE Trans. Systems Man Cybernet. 22 (4), 688–704.

Bhatnagar, R.K., Kanal, L.N., 1986. Handling uncertain information: A review of numeric and non-numeric methods. In: Kanal, L.N., Lemmer, J.F. (Eds.), Uncertainity in Artificial Intelligence. Elsevier Science Publishers, pp. 3–26.

Bhattacharya, P., 2000. On the Dempster–Shafer evidence theory and non-hierarchical aggregation of belief structures. IEEE Trans. Systems Man Cybernet. 30 (5), 526–536.

Bloch, I., 1996. Information combination operators for data fusion: A comparative review with classification. IEEE Trans. Systems Man Cybernet. 26 (1), 52–67.

Brunelli, R., Falavigna, D., 1995. Person identification using multiple cues. IEEE Trans. Pattern Anal. Machine Intell. 17 (10), 955–966.

Campbell, J.P., 1997. Speaker recognition: A tutorial. Proc. IEEE 85 (9), 1437–1462.

Chen, K., Chi, H., 1998. A method of combining multiple probabilistic classifiers through soft competition on different feature sets. Neural Computing 20, 227–252.

Chen, K., Wang, L., Chi, H., 1997. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. Internat. J. Pattern Recognition Artificial Intell. 11 (3), 417–445.

Chen, K., Xie, D., Chi, H., 1996. Combine multiple time-delay HMEs for speaker identification. In: IEEE International Conference on Neural Networks, Vol. 4, pp. 2015–2020.

Doddington, G.R. et al., 2000. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. Speech Commun. 31, 225–254.

Farrell, K.R., 1995. Text independent speaker verification using data fusion. In: IEEE-ICASSP Proceedings, pp. 349–352.

Farrell, K.R., Mammone, R.J., 1995. Data fusion techniques for speaker recognition. In: Ramachandran, R., Mammone, R.J. (Eds.), Modern Methods of Speech Processing. Kluwer Academic Publishers, pp. 279–297 (Chapter 12).

Farrell, K.R., Ramachandran, R.P., Mammone, R.J., 1998. An analysis of data fusion methods for speaker verification. In: IEEE-ICASSP Proceedings, Vol. 2, pp. 1129–1132.

Fredouille, C., Bonastre, J.F., Merlin, T., 2000. AMIRAL: A block-segmental multirecognizer architecture for automatic speaker recognition. Digital Signal Process. 10, 172–197.

Fung, R.M., Chong, C.Y., 1986a. Metaprobability and Dempster–Shafer in evidental reasoning. In: Kanal, L.N., Lemmer, J.F. (Eds.), Uncertainity in Artificial Intelligence. Elsevier Science Publishers, pp. 295–303.

Fung, R.M., Chong, C.Y., 1986b. Confidence factors, empiricism and the Dempster–Shafer theory of evidence. In: Kanal, L.N., Lemmer, J.F. (Eds.), Uncertainity in Artificial Intelligence. Elsevier Science Publishers, pp. 117–125.

Furui, S., 1997. Recent advances in speaker recognition. Pattern Recognition Lett. 18 (9), 859–872.

Genoud, D. et al. 1996. Combining methods to improve speaker verification decision. In: ICSLP Proceedings, Vol. 3, pp. 1756–1759.

Gish, H., Schmidt, M., 1994. Text-independent speaker identification. IEEE Signal Process. Mag., 18–32.

Heck, L.P. et al., 2000. Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Commun. 31, 181–192.

Heck, L.P., Weintraub, M., 1997. Handset-dependent background models for connected digit password speaker verification. In: IEEE-ICASSP Proceedings, April.

Hegarat-Mascle, S.L., Bloch, I., Vidal-Madjar, D., 1998. Introduction of neighborhood information in evidence theory and application to data fusion of radar and optical images with partial cloud cover. Pattern Recognition 31 (11), 1811–1823.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Acoustics, Speech Signal Process. 2 (4), 578–589.

Hermansky, H., Morgan, N., Bayya, A., Kohn, P., 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In: EUROSPEECH Proceedings.

Ho, T.K., Hull, J., Srihari, S., 1994. Decision combination in multiple classifier systems. IEEE Trans. Pattern Anal. Machine Intell. 16, 66–75.

Melin, H., Lindeberg, J., 1997. Guidelines for experiments on the POLYCOST database, January.

Pigeon, S., Druyts, P., Verlinde, P., 2000. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. Digital Signal Process. 10, 237–248.

Quatieri, T.F., 2002. Discrete-Time Speech Signal Processing PRINCIPLES AND PRACTICE. Prentice Hall Inc.

Quatieri, T.F., Reynolds, D.A., O'Leary, G., 1998. Magnitude-only estimation of handset nonlinearity with applications to speaker recognition, In: IEEE-ICASSP Proceedings.

Radova, V., Psutka, J., 1997. An approach to speaker identification using multiple classifiers. In: IEEE-ICASSP Proceedings, pp. 1135–1138.

Rahman, A.F.R., Fairhurst, M.C., 1999. Enhancing multiple expert decision combination strategies through exploitation of *a priori* information sources. IEE Proc.-Vis. Image Signal Process. 146, 40–49.

Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker recognition using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3 (1), 72–83.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10, 19–41.

Reynolds, D.A., 1997. Comparison of background normalization methods for text-independent speaker verification. In: EUROSPEECH Proceedings.

Shafer, G., Logan, R., 1987. Implementing Dempster's rule for hierarchical evidence. Artificial Intell. 33, 271–298.

Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press.

Smets, P., Kennes, R., 1994. The transferrable belief model. Artificial Intell. 66, 191–234.

Soong, F.K., Rosenberg, A.E., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoustics, Speech Signal Process. 36 (6), 871–879.

Voorbraak, F., 1991. On the justification of Dempster's rule of combination. Artificial Intell. 48, 171–197.

Xu, L., Krzyzak, A., Suen, C.Y., 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Systems Man Cybernet. 22, 418–435.