

# A NEW ALGORITHM FOR SPEAKER IDENTIFICATION USING THE DEMPSTER-SHAFER THEORY OF EVIDENCE

Imran Naseem and Mohamed Deriche (*inaseem, mderiche@kfupm.edu.sa*)

Electrical Engineering Department,  
King Fahd University of Petroleum and Minerals,  
Dhahran 31261, Saudi Arabia.

## ABSTRACT

In this paper, speaker identification using the Dempster-Shafer theory of evidence is discussed. The objective is to use the complementary information present from different classifiers to fuse the classification results into a single decision. Here, we use a decreasing function of the distance (of the classifiers) as our belief function. In the case of speaker identification, we show that a combined classifier based on the Dempster-Shafer theory outperforms the individual LPCC and MFCC classifiers when used separately.

## 1. INTRODUCTION

The ultimate goal of robust pattern recognition systems is to achieve the classification accuracy for the task at hand. This objective traditionally led to the development of different classification schemes for any pattern recognition problem to be solved. However various classifiers differ from each other in terms of the different feature sets used and therefore exhibit complementary information [1]. It follows that combining different classifiers for a given problem in an *efficient* way would result in an improved classification rate [2]. In this context Dempster-Shafer Theory of Evidence (DST) [3] has shown some promising results [4], [2]. The DST [3] is a powerful tool for representing uncertain knowledge. This theory has inspired many researchers to investigate different aspects related to uncertainty and lack of knowledge and their applications to real life problems [5]. Classifier combination is one of such problems. In this paper, we have developed a generic algorithm, the NNEF (Nearest Neighbor based Evidence Fusion), to combine nearest neighbor classifiers using the DST. Here the algorithm is applied to the problem of speaker recognition.

With the exponential growth in technology and the growth in business carried worldwide, it is becoming crucial to build automated systems that identify people. Traditionally, body characteristics such as face and voice have been successfully used in identification. Such body characteristics or more generally biometrics, first came into extensive use for

law-enforcement and legal purposes, identification of criminals and illegal aliens. It was expanded for usage in security clearances for employees in sensitive jobs, paternity determinations, forensics, positive identifications of convicts and prisoners, and so on. Today, however, many civilian and private-sector applications are increasingly using biometrics to establish personal identification. Nowadays commonly used biometrics include: fingerprints, hand or palm geometry, retina, iris and facial characteristics. Behavioral characters include signature, voice (which also has a physical component), keystroke pattern, and gait. Of this class of biometrics, technologies for face and voice are the most established. Voice biometric has the most potential for growth, because it requires no new hardware, as most PCs already contain a microphone. However, poor quality microphones and ambient noise can severely affect verification.

## 2. COMBINING CLASSIFIERS USING THE DEMPSTER-SHAFER THEORY OF EVIDENCE

### 2.1. Motivation and Fundamental Concepts

Let us first give a simple example to explain the concept uncertainty. Let  $\theta$  represent the following proposition: the *passionfruit* is delicious. Then, according to the Bayesian theorem,  $P(\theta) + P(\bar{\theta}) = 1$ , where  $\bar{\theta}$  is negation of  $\theta$ . Now suppose that Jim has not tasted the passionfruit before. Then, we cannot say that Jim believes the proposition if he has no idea what it means. Also, it is not fair to say that he disbelieves the proposition. This problem can be better represented by the Dempster-Shafer (D-S) theory of evidence, which is regarded as a more general approach to representing uncertainty than the Bayesian approach. The D-S theory would denote Jim's belief of the proposition,  $m(\theta)$ , and disbelief,  $m(\bar{\theta})$ , as both being zero. The Bayesian approach does not allow this.

Thus, the difference between the Bayesian statistical model and the D-S evidential theory is conceptual. In the statistical model, it is assumed that there is a Boolean phenomena which either does or does not exist. The result of this as-

sumption leads to the implication that commitment of belief to a hypothesis leads to the commitment of the remaining belief to its negation. If there is little belief for the existence of a phenomena this would imply, under the Bayesian formulation, a large belief to its non-existence, which is what we call *over-commitment*. In D-S theory, one considers the evidence in favor of a hypothesis. There is no causal relationship between a hypothesis and its negation, hence lack of belief does not imply disbelief. Rather, lack of belief in any particular hypothesis implies belief in the set of all hypotheses, which is referred to as the state of uncertainty. If we denote the uncertainty by  $\Theta$ , then in the above example  $m(\Theta) = 1$ , which is calculated by the following formula:  $m(\theta) + m(\bar{\theta}) + m(\Theta) = 1$ . In this work, we will use such D-S theory to solve the problem of speaker identification.

## 2.2. The Dempster-Shafer Theory of Evidence

The D-S theory of evidence [3] is a powerful tool for representing uncertain knowledge. This theory has inspired many researchers to investigate different aspects related to uncertainty and lack of knowledge and their applications to real life problems [5]. Today, the D-S theory covers several different models including the transferable belief model (TBM) [3].

In order to explain the combination rule under the TBM model, we need to present the definitions of basic belief assignment and belief function. Let  $\Theta = \{\theta_1, \dots, \theta_K\}$  be a finite set of possible hypotheses. This set is referred to as the frame of discernment, and its powerset denoted by  $2^\Theta$ . The basic belief assignment of a subset of  $\Theta$  and the belief function associated with it are defined as follows:

### 2.2.1. Basic Belief Assignment (BBA)

A basic belief assignment  $m(\cdot)$  is a function that assigns a value in  $[0,1]$  to every subset  $\mathbf{A}$  of  $\Theta$  and satisfies the following:

$$m(\phi) = 0, \text{ and } \sum_{\mathbf{A} \subseteq \Theta} m(\mathbf{A}) = 1 \quad (1)$$

where  $\phi$  is the empty set. It is worth noting that  $m(\phi)$  can be non-zero when considering un-normalized combination rule as will be explained later. While in probability theory a measure of probability is assigned to atomic hypotheses  $\theta_i$ ,  $m(\mathbf{A})$  is the measure of belief that supports  $\mathbf{A}$ , but does not support anything more specific, i.e., strict subsets of  $\mathbf{A}$ . For  $\mathbf{A} \neq \theta_i$ ,  $m(\mathbf{A})$  reflects some ignorance because it is a belief that we cannot subdivide  $\mathbf{A}$  into finer subsets.  $m(\mathbf{A})$  is a measure of support we are willing to assign to a composite hypothesis  $\mathbf{A}$  at the expense of support  $m(\theta_i)$  of atomic hypotheses  $\theta_i$ . For a particular frame of discernment  $\Theta$ , if we set  $m(\theta_i \neq 0)$  for all  $\theta_i$  and  $m(\mathbf{A}) = 0$  for all  $\mathbf{A} \neq \theta_i$ , then  $m(\theta_i)$  becomes probability of  $\theta_i$  with

$\sum_i m(\theta_i) = 1$ . A subset  $\mathbf{A}$  for which  $m(\mathbf{A}) > 0$  is called a focal element. The partial ignorance associated with  $\mathbf{A}$  leads to the following inequality:  $m(\mathbf{A}) + m(\bar{\mathbf{A}}) \leq 1$ , where  $\bar{\mathbf{A}}$  is the complement of  $\mathbf{A}$ . In other words, the D-S theory of evidence allows us to represent only our actual knowledge without being forced to *overcommit* when we are ignorant.

### 2.2.2. Belief Function

The belief function,  $bel(\cdot)$ , associated with the BBA  $m(\cdot)$  is a function that assigns a value in  $[0,1]$  to every nonempty subset  $\mathbf{B}$  of  $\Theta$ . It is called *degree of belief in  $\mathbf{B}$*  and is defined by

$$bel(\mathbf{B}) = \sum_{\mathbf{A} \subseteq \mathbf{B}} m(\mathbf{A}) \quad (2)$$

where  $\mathbf{A}$  is subset of  $\mathbf{B}$ . We can consider a basic belief assignment as a generalization of a probability density function whereas a belief function is a generalization of a probability distribution function.

### 2.2.3. Combination rule

Consider two BBAs  $m_1(\cdot)$  and  $m_2(\cdot)$  for belief functions  $bel_1(\cdot)$  and  $bel_2(\cdot)$  respectively. Let  $\mathbf{A}_j$  and  $\mathbf{B}_k$  be focal elements of  $bel_1(\cdot)$  and  $bel_2(\cdot)$  respectively. Then  $m_1(\cdot)$  and  $m_2(\cdot)$  can be combined to obtain the belief committed to  $\mathbf{C} \subseteq \Theta$  according to the following combination or *orthogonal sum* formula [3],

$$\begin{aligned} m(\mathbf{C}) &= m_1(\mathbf{C}) \oplus m_2(\mathbf{C}) \quad (3) \\ &= \frac{\sum_{j,k, \mathbf{A}_j \cap \mathbf{B}_k = \mathbf{C}} m_1(\mathbf{A}_j) m_2(\mathbf{B}_k)}{1 - \sum_{j,k, \mathbf{A}_j \cap \mathbf{B}_k = \phi} m_1(\mathbf{A}_j) m_2(\mathbf{B}_k)}, \mathbf{C} \neq \phi \end{aligned}$$

The denominator is a normalizing factor, which intuitively measures how much  $m_1(\cdot)$  and  $m_2(\cdot)$  are conflicting. Smets [6] proposed the un-normalized combination rule :

$$m_1(\mathbf{C}) \cap m_2(\mathbf{C}) = \sum_{j,k, \mathbf{A}_j \cap \mathbf{B}_k = \mathbf{C}} m_1(\mathbf{A}_j) m_2(\mathbf{B}_k), \quad (4) \quad \forall \mathbf{C} \subseteq \Theta$$

Here we propose to use the D-S theory to the problem of combining the classification results of different classifiers by considering the evidence of each classifier as a BBA. Since the classifiers' evidence plays a crucial role in the combination performance, there is an increased interest in the proper estimation of such evidence. In the next section, we discuss our proposed method of estimating belief in a distance classifier and its implementation for the speaker recognition problem.

### 3. THE PROPOSED DST FRAMEWORK FOR CLASSIFIER COMBINATION

#### 3.1. Dempster-Shafer Formulation of the Problem

Consider the case of  $N$  classifiers denoted by  $e^{(n)}$ , where  $n = 1, 2, \dots, N$ . Let  $\mathbf{X}_k$  be the training data matrix for each class,  $k = 1, 2, \dots, K$ ,  $K$  being total number of classes. We will assume here equal amount of training for each of the classes. Also let  $\theta_k$  be the label for each class  $k$ . Now, the feature extraction module of each classifier extracts a feature matrix  $\mathbf{X}_k^{(n)}$ . We define a modeling function  $\Omega(\cdot)$  which models each class so that

$$\Omega(\mathbf{X}_k^{(n)}) = \mathbf{U}_k^{(n)}; \quad k = 1, 2, \dots, K \quad (5)$$

$$n = 1, 2, \dots, N \quad (6)$$

Let  $\mathbf{z}$  be an input test pattern which is modeled in a similar way :

$$\Omega(\mathbf{z}) = \mathbf{Z} \quad (7)$$

For the case of a single classifier, the classification task is to assign class  $i$  to pattern  $\mathbf{z}$  if:

$$D(\mathbf{U}_i, \mathbf{Z}) < D(\mathbf{U}_k, \mathbf{Z}) \quad \forall k = 1, 2, \dots, K \quad (k \neq i) \quad (8)$$

where  $\mathbf{U}_k$  is the model for each class  $k$ , and  $\mathbf{U}_i$  being the nearest neighbor to  $\mathbf{Z}$ .  $D(\cdot)$  is a distance measure between the test pattern model ( $\mathbf{Z}$ ) and the training pattern models for each class ( $\mathbf{U}_k \quad k = 1, 2, \dots, K$ ).

Assume now that we have  $N$  classifiers, so that each classifier operates on the test model independently to reach an independent decision.

Since for each classifier, the function  $\Omega(\cdot)$  models the patterns in the same manner, we propose the nearest neighbor distance  $\underbrace{\min}_k^{(n)} \{D(\mathbf{U}_k^{(n)}, \mathbf{Z})\}$  as the evidence of our belief in the decision made by classifier  $n$ . Thus, the belief becomes a decreasing function (say  $\psi(\cdot)$ ) of this distance:

$$m^{(n)}(i) = \psi(\underbrace{\min}_k^{(n)} \{D(\mathbf{U}_k, \mathbf{Z})\}) \quad (9)$$

where  $m^{(n)}(i)$  is our belief in classifier  $n$  for classifying test pattern  $\mathbf{z}$  as class  $i$ .

One candidate for the function  $\psi(\cdot)$  could be the exponential function:

$$m^{(n)}(i) = \exp(-(\underbrace{\min}_k^{(n)} \{D(\mathbf{U}_k, \mathbf{Z})\})) \quad (10)$$

Hence the smaller the nearest neighbor distance measure, the greater is our belief in the decision of the classifier. In summary our algorithm works as follows:

1. Each class is modeled using the training data matrix  $\mathbf{X}_k, k = 1, 2, \dots, K$  and the function  $\Omega(\mathbf{X}_k^{(n)}) = \mathbf{U}_k^{(n)}$ .
2. Input test pattern  $\mathbf{z}$  is also modeled using the same modeling function  $\Omega(\cdot)$ , i.e  $\Omega(\mathbf{z}) = \mathbf{Z}$ .
3. A distance measure,  $D(\cdot)$  is then used to evaluate the distance between  $\mathbf{Z}$  and each of the models  $\mathbf{U}_k^{(n)}, k = 1, 2, \dots, K$ .
4. For each classifier, a label is given to the test pattern  $\mathbf{z}$  which corresponds to minimum distance measure

$$d^{(n)} = \underbrace{\min}_k \{D(\mathbf{U}_k^{(n)}, \mathbf{Z}^{(n)})\} \quad (11)$$

$$n = 1, 2, \dots, N$$

$$k = 1, 2, \dots, K$$

5. We estimate our confidence in each classifier's decision as:

$$m^{(n)}(i) = \exp(-d^{(n)}) \quad (12)$$

6. We then combine all evidences  $m^{(n)} \quad n = 1, 2, \dots, N$  using Dempster-Shafer theory of evidence as follows:

$$m(k) = \frac{\sum_{j,l, \mathbf{A}_j \cap \mathbf{A}_l = k} m^{(1)}(\mathbf{A}_j) \dots m^{(N)}(\mathbf{A}_l)}{1 - \sum_{j,l, \mathbf{A}_j \cap \mathbf{A}_l = \phi} m^{(1)}(\mathbf{A}_j) \dots m^{(N)}(\mathbf{A}_l)} \quad (13)$$

$$k = 1, 2, \dots, K$$

7. Class label  $j$  is assigned to test pattern if

$$j = \underbrace{\max}_k \{m(k)\}; \quad k = 1, 2, \dots, K \quad (14)$$

Some special cases to be considered are:

- (a) if all classifiers reject a pattern, the consensus decision will then be rejection and thus our belief will be given to the frame of discernment  $m(\Theta) = 1$ .
- (b) if a subset of classifiers say  $M$  rejects a test pattern, then these classifiers will be excluded and the decision will be made on basis of remaining  $(N - M)$  classifiers.

#### 4. THE DEVELOPED SPEAKER RECOGNITION SYSTEM

Among the biometric traits available for the purpose of person identification, speech makes the most natural and obvious choice. Automatic speaker recognition (ASR) systems identify people utilizing the utterances.

We have developed two different speaker recognition systems, the main difference between the two systems resides in the different features used. Specifically we used the LPCC (Linear Prediction Cepstral Coefficients) and MFCC (Mel Frequency Cepstral Coefficients) methods of feature extraction.

##### 4.1. Feature Extraction through LPCC

One of the most popular speech analysis techniques is that of linear prediction. The basic idea behind linear prediction analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. The mathematical details could be found in [7]. We have used the work in [8] to implement an LPCC system. The order of the linear prediction filter is 10.

##### 4.2. Feature Extraction through MFCC

MFCC is perhaps the best known and the most popular feature extraction technique for speech signals. The main purpose of the MFCC is to imitate the behavior of a human ear. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency  $f$ , measured in hertz, a subjective pitch is measured on a scale called the *Mel* scale [9]. We have used the work in [10] to obtain the MFCC features.

##### 4.3. The Speaker Recognition System

We consider second order statistical modeling of speech, assuming a wide sense stationary process (WSS) as proposed in [11]. Let  $\mathbf{X}_k$  be the training data matrix for class  $k$ , so that we have  $b$  samples available per class for training. Let  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_b\}$ , be a set of  $b$  feature vectors available for class  $k$ . Given the  $b$  patterns available for training per class, we model a class  $k$  through the mean and covariance matrix as follows:

$$\hat{\mathbf{x}} = \frac{1}{b} \sum_{i=1}^b \mathbf{x}_i \quad (15)$$

$$(16)$$

$$\mathbf{U}_k = \frac{1}{b} \sum_{i=1}^b (\mathbf{x}_i - \hat{\mathbf{x}})(\mathbf{x}_i - \hat{\mathbf{x}})^t \quad (17)$$

Similarly for a test pattern  $\mathbf{z}$ , we derive a covariance matrix  $\mathbf{Z}$ . Once we have developed the second-order statistical models, we apply an *arithmetic-harmonic sphericity* measure [11] as the distance metric between  $\mathbf{X}$  and  $\mathbf{Z}$ , thus

$$D_{sph}(\mathbf{U}_k, \mathbf{Z}) = \log \left[ \frac{\text{tr}(\mathbf{U}_k \mathbf{Z}^{-1}) \text{tr}(\mathbf{Z} \mathbf{U}_k^{-1})}{m^2} \right];$$

$$k = 1, 2, \dots, K. \quad (18)$$

where  $m$  is the dimension of the feature vector and  $\text{tr}(\mathbf{A})$  is trace of  $\mathbf{A}$ . The distance measures are mapped to  $[0, 1]$  with a sigmoid function.

##### 4.4. DST based Fusion of Speaker Recognition Systems using the Proposed NNEF Algorithm

We are now at the stage of testing our proposed fusion algorithm. Note that although the features are heterogeneous, they are reduced to the same distance metric, and thus we can safely take the distance as an evidence measure since there is no need for normalization. We have used a locally developed text-dependent database consisting of 40 classes. The password for authentication is the arabic greeting sentence *assalam-o-alaikum wa rahmatullah-e-wabarakathu*.

We tested our algorithm under three evaluation protocols:

1. **Evaluation Protocol 1:** Under the first evaluation protocol we verify our fusion algorithm (NNEF) when we assume one classifier is perfect. In our case it is the MFCC based classifier, which resulted in a 0% rejection and substitution (misclassification) rate. The NNEF algorithm also resulted in a 100% classification accuracy. Thus the evidence of MFCC classifier is strong enough to dominate the decision of LPCC, the DST fusion of the two, thereby giving optimal results.

Method	Training	Testing	Recog.
MFCC	5	3	100%
LPCC	5	3	91.67%
NNEF	5	3	100%

**Table 1.** DST fusion results under evaluation protocol 1

Table 1 clearly shows that when combining a perfect classifier with a poor classifier, the proposed NNEF algorithm opts for the perfect classifier, thereby avoiding the averaging phenomenon.

2. **Evaluation Protocol 2:** Under the second evaluation protocol, we modify our MFCC and LPCC classifiers by introducing a threshold value  $\alpha$  for rejection. The

Method	Training	Testing	Recog.
MFCC	5	3	91.67%
LPCC	5	3	87.5%
NNEF	5	3	95.83%

**Table 2.** DST fusion results under evaluation protocol 2

values of  $\alpha$  being 0.53 and 0.6 for MFCC and LPCC classifiers respectively.

The recognition accuracy of the two classifiers has thus been reduced to 91.67% for MFCC and 87.5% for LPCC (see table 2). The DST based fusion of the two classifiers' decision using the proposed NNEF algorithm outperformed the two individual classifiers giving an improved recognition rate of 95.83%.

3. **Evaluation Protocol 3:** Under Evaluation Protocol 3, we make the problem more complicated by adding white Gaussian noise to the speech data. The aim is to verify the robustness of the NNEF algorithm under noisy conditions.

Method	Training	Testing	Recog.
MFCC	5	3	87.5%
LPCC	5	3	86.6%
NNEF	5	3	93.33%

**Table 3.** Results of the NNEF algorithm for 20dB SNR

The results for noise contaminated speech data for different SNR values are shown in tables 3 and 4. For 20dB and 15dB SNR the NNEF algorithm shows an improvement of 6% and 5% respectively, over the best of the combining classifiers.

## 5. CONCLUSION

In this paper, we developed a new approach for combining the results of 2 different classifiers used in speaker identification. In particular, we use the concept of evidence to combine two or more different classifiers. Based on such concept, we showed that we can improve the classification results of the LPCC and the MFCC classifiers when used separately.

Method	Training	Testing	Recog.
MFCC	5	3	85%
LPCC	5	3	83.3%
NNEF	5	3	90%

**Table 4.** Results of the NNEF algorithm for 15dB SNR

## 6. ACKNOWLEDGEMENT

The authors would like to thank King Fahd University of Petroleum and Minerals for the support provided to carry this research.

## 7. REFERENCES

- [1] R.P.W.Duin J.Kittler, M.Hatef and J.Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.
- [2] A. Krzyzak L. Xu and C.Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, pp. 418–435, 1992.
- [3] G. Shafer, "A mathematical theory of evidence," *Princeton University Press*, 1976.
- [4] G. Rogova, "Combining the results of several neural network classifier," *Neural Networks*, vol. 7, pp. 777–781, 1994.
- [5] I. Bloch N. Milisavljevic and M. Acheroy, "Characterization of mine detection sensors in terms of belief functions and their fusion, first results," *In 3rd Intl. Conf. on Information Fusion*, 2000.
- [6] P. Smets, "The transferable belief model for quantified belief representation," *In D.M. Gabbay and P. Smets, editors, Handbook of defeasible reasoning and uncertainty*, vol. Kluwer, pp. 267–301, 1998.
- [7] Wilpon J.G. Ephraim, Y. and L.R. Rabiner, "A linear predictive front-end processor for speech recognition in noisy environments," *IEEE Transactions ASSP*, 1987.
- [8] B.S.Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *JASA*, pp. 1304–1312, 1974.
- [9] B.C.J.Moore, *Frequency Analysis and Masking*, Academic Press, USA, 1995.
- [10] A.Acero X.Haung and H.W.Hon, *Spoken Language Processing: A guide to theory, algorithm and system development*, Prentice Hall PTR, New Jersey,, 2001.
- [11] I.Magrín-Chagnolleau F.Bimbot and L.Mathan, "Second-order statistical measure for text-independent speaker identification," *Speech Communication*, vol. 17, pp. 177–192, 1995.