COMPARISON OF DIFFERENT OBJECTIVE FUNCTIONS FOR OPTIMAL LINEAR COMBINATION OF CLASSIFIERS FOR SPEAKER IDENTIFICATION

Hakan Altınçay

Computer Engineering Dept.
Eastern Mediterranean University
Gazi Mağusa, North Cyprus
E-mail: hakan.altincay@emu.edu.tr

Mübeccel Demirekler

Electrical and Electronics Engineering Dept.
Middle East Technical University
Ankara, Turkey
E-mail: demirek@metu.edu.tr

ABSTRACT

This paper presents a comparison of objective functions for optimally combining different speaker identification systems. The comparison is based on the classification performance of the resultant multiple classifier system (MCS). The objective functions considered are; classification figure of merit (CFM), mean square error (MSE) and cross entropy (CE). In all three methods, the outputs of individual classifiers assumed to be the posterior probabilities of each speaker and linear combination of the output vectors are considered. CFM seeks to maximize the difference between the output value of the speaker and the output values of all other incorrect speakers. On the other hand, MSE and CE compares the outputs with some ideal vectors where the output of the correct speaker is set to one and the others are zero. The experimental results are also compared with the averaging method where the combination is not optimized. Our simulation experiments on four different sets of speakers have shown that CFM performs better compared to the other objective functions.

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 architecture 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 performance of classification systems. There have been extensive research in this field and promising results were obtained [1].

Classification figure of merit (CFM) was proposed in [2] for training time-delay neural network weights for phoneme recognition and later used in [3] as a *ranking figure of merit* which rewards the classifiers to obtain better ranking. Ueda also used this objective function (OF) to optimally combine several neural networks [4]. CFM in our case serves the purpose of finding optimal weights for individual classifiers in a combination scheme. The classifiers in this study are assumed to produce posterior probabilities for each speaker for a given test utterance. During combination, posterior probability vectors produced by each classifier are linearly combined by selecting optimum weights according to CFM function. The entries of the combined vector are called the outputs of the system. CFM seeks to maximize the difference between the output value for the correct speaker and the output values of all other incorrect speakers. It has some potential differences compared to other OFs. Firstly, it has no notion of an ideal target classification output pattern to which it should match its output. Output value of the correct speaker is only forced to be larger than the output value of any other speaker. Secondly the objective function saturates for large differences between the output values of the correct and incorrect speakers, meaning, no effort has been made to identify a speaker ideally but instead it is tried to identify more speakers with non-ideal outputs. In other words it is not necessary that combined probability of the correct speaker is close to 1 or even much higher that the probabilities of all other speakers but it is only required to be somewhat larger than the others.

MSE and CE [2] are well known OFs that are more frequently used for optimal classifier combination. They compare the outputs with some ideal vectors where, in selecting the ideal vectors, the output of the correct speaker is set to one and the others are zero. As an example, Hashem studied the optimal classifier combination (OCC) problem using MSE criterion [1].

In this study, these OFs are compared for optimal combination of classifiers for speaker identification problem. Some simulation experiments are conducted and their relative performances are analyzed in terms of correct classification and better rankings and then, the results are compared with averaging method where the combination is not optimized. Instead, the output vectors are added and the speaker getting the largest combined output is selected as the joint decision.

2. NOTATION

The combination scheme assumes that the output of each of the classifier is a vector where ith entry is the probability that the tested input pattern data belongs to speaker c_i. Let $p_m(c_i|e_m(t))$ denote the a posteriori probability that the tested input pattern data belongs to speaker c_i given the output of the classifier $e_m(.)$ where t denotes a speech token and m denotes the classifier. Define $P_m(t)$ as the vector of a posteriori probabilities of all speakers. Let W_m be an NxN diagonal matrix which denotes the weights of the mth classifier.

$$W_m = \begin{bmatrix} w_{m,1} & & & \\ & w_{m,2} & O & \\ & O & \ddots & \\ & & & w_{m,N} \end{bmatrix} \quad (1)$$

Assume that we have N speakers and M classifiers. The a posteriori probability of all speakers after combination is denoted by the vector $\overline{O}(t) = \sum_{m=1}^{M} W_m.P_m(t)$ where the ith element of the vector, $O_i(t)$, gives the combined a posteriori probability of speaker c_i and t denotes the speech token. As an example, for the case of 2 classifiers, the combined output $\overline{O}(t)$ is;

$$\overline{O}(t) = \begin{bmatrix} w_{1,1}p_1(c_1|e_1(t)) + w_{2,1}p_2(c_1|e_2(t)) \\ w_{1,2}p_1(c_2|e_1(t)) + w_{2,2}p_2(c_2|e_2(t)) \\ \vdots \\ w_{1,i}p_1(c_i|e_1(t)) + w_{2,i}p_2(c_i|e_2(t)) \\ \vdots \\ w_{1,N}p_1(c_N|e_1(t)) + w_{2,N}p_2(c_N|e_2(t)) \end{bmatrix} \quad (2)$$

Notice that, for averaging rule, the weight matrices are selected as the identity matrices.

3. CLASSIFICATION FIGURE OF MERIT OBJECTIVE FUNCTION, CFM

CFM is used to maximize the *difference* in the a posteriori probability between the correct speaker and *all* of the other (incorrect) speakers, i.e. $O_c(t) - O_n(t)$ where $O_c(t)$ is the combined output for the correct speaker and $O_n(t)$ is the combined output for nth speaker. While doing this, *CFM* does not aim at maintaining large differences between these a posteriori probabilities. Instead, it rewards for decreased misclassifications. In other words, for *CFM*, the weights that give less number of misclassification is closer to optimal solution. This aim requires an objective function which saturates at a certain level when $O_c(t) - O_n(t)$ gets larger values. One very simple function which satisfies this requirement is the unit step function. Such a function may perform very well for validation data, but may not give a robust result since, even small differences between $O_c(t)$ and $O_n(t)$ are rewarded. The sigmoid function with a controlled slope at $O_c(t) = O_n(t)$ is considered to be more suitable and proposed in [2].

There are several different forms for this *OF*. In this study, we considered first and Nth order *CFM*.

3.1. *N*th Order CFM Objective Function

The objective function for the Nth order *CFM* is defined as

$$CFM = \frac{1}{N-1} \sum_{t=1}^{T} \sum_{\substack{n=1 \\ n \neq c}}^{N} (1 + exp(-\beta \delta_n(t)))^{-1} \quad (3)$$

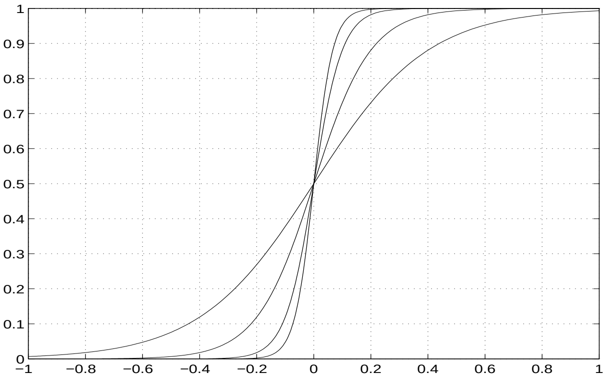

Figure 1: The sigmoid function for different β values.

where $\delta_n(t) = O_c(t) - O_n(t)$. $O_c(t)$ is the a posteriori probability of the correct speaker and $O_n(t)$ is the a posteriori probability of speaker c_n. β is the sigmoid discontinuity parameter. t denotes the validation tokens used for optimal weight estimation. For $\beta = 5, 10, 20, 30$, the sigmoid function is given in Figure 1. For small values of β, the function is more flat and rewards for *larger differences* between the correct speaker and other speakers. Actually, from classification point of view, this is not that much important. Instead, we seek only for positive values for $\delta_n(t)$. This dictates the use of very large values of β such that the function approaches to the unit step. But, increase in β yields increasingly discontinuous objective function which results in slow and unstable searches [2]. Furthermore, such a choice of β may introduce some robustness problems. Experimental results for different β values will be given in section 6. Our experiments have shown that $\beta = 20$ is a suitable choice.

We aim at maximizing the CFM objective function. This is done using the steepest ascent (not descent) algorithm as follows.

$$W_m(k+1) = W_m(k) + \alpha \frac{\partial CFM}{\partial W_m(k)} \quad (4)$$

3.2. 1st Order CFM ($CFM_{\delta_{min}}$)

This is a special case for the the general Nth order CFM where the difference between the a posteriori probability of the correct speaker and the maximum a posteriori probability among incorrect speakers is minimized. The objective function for this case is defined as:

$$CFM = \frac{1}{N-1} \sum_{t=1}^{T} (1 + \exp(-\beta \delta_{min}(t)))^{-1}, \quad (5)$$

and,

$$\delta_{min} = \min_{\substack{n \\ n \neq c}} (O_c(t) - O_n(t)). \quad (6)$$

4. MEAN SQUARE ERROR (MSE) OBJECTIVE FUNCTION

The MSE function aims at minimizing the mean squared error between the ideal or desired a posteriori probability and the joint a posteriori probability of the classifiers for all classes. The objective function is defined as:

$$MSE = \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} (O_n(t) - D_n(t))^2 \qquad (7)$$

where $O_n(t)$ denotes the output probability of speaker c_n of the combined classification system and $D_n(t)$ is the desired output for the corresponding speaker. $D_n(t) = 1$ when $n = c$ and zero for other speakers. We aim at minimizing the MSE objective function. This is done using the steepest descent algorithm,

$$W_m(k+1) = W_m(k) - \alpha \frac{\partial MSE}{\partial W_m(k)}. \qquad (8)$$

5. CROSS ENTROPY (CE) OBJECTIVE FUNCTION

The CE objective function seeks to minimize the difference between the ideal a posteriori probability and the joint a posteriori probability of the classifiers for all speakers by minimizing the cross entropy between the actual and desired outputs. The objective function is defined as [3]:

$$CE = -\frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \Big[D_n(t) \log(O_n(t)) + (1 - D_n(t)) \log(1 - O_n(t)) \Big]. \qquad (9)$$

The optimal weight estimation is done using the steepest descent algorithm. The desired output function is the same as the one for MSE case.

6. SIMULATION EXPERIMENTS

The simulation experiments are conducted on the POLYCOST database [5]. This database contains telephone speech sampled at 8kHz. On the average, 10 sessions are recorded by each of 74 male and 60 female speakers from 14 different countries. Each session contains 20 seconds of speech on the average. In the simulations, 4 different sets of male speakers, SET1,...,SET4 are used. The speakers in these sets are given in Table 1. Due to insufficiency in training or validation data, the files in m042, m045 and m058 are replaced by m061, m062 and m063. First three sessions of the mother tongue records that contain free text speech are used both for training and cross validation. Class probabilities are obtained during cross validation using 20 tokens from each session. Two records starting from session five are used for testing where, each of these records is partitioned into 10 tokens. Each token is treated as a different session.

Test sets	Speakers included
SET1	m001,...,m030
SET2	m031,...,m060
SET3	m015,...,m029,m031,...,m045
SET4	m001,...,m015,m046,...,m060

Table 1: The speakers included in different speaker sets.

Classifiers	Test data	Validation data
classifier 1	80.58%	84.55%
classifier 2	84.97%	86.14%
Averaging	83.69%	90.66%
MSE	87.11%	90.61%
CE	87.92%	91.00%
CFM	89.25%	94.28%
$CFM_{\delta_{min}}$	87.80%	93.11%

Table 2: Identification rates of different classification systems on SET1.

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 [6]. For the first classifier, $e_1(.)$, cepstral mean subtraction 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 models are used for the modeling of the speakers [7]. These classifiers were shown to provide complementary information for each other in [8].

The experimental results of the given optimal weight estimation methods using the test tokens on SET1 are given in Table 2. The results for other sets are given in Tables 3-5. The performance of CFM is better than the other objective functions when 'first speaker' or 'first two speakers' or 'first three speakers' are selected as the output of the system.

Ranking performances of different OCC schemes on validation tokens are given in Figure 2. The figure gives the ranking performance of the CFM objective function for three different values of β on the validation sessions of SET1. As seen from the figure, with less than 1.0% er-

Classifiers	Test data	Validation data
classifier 1	71.29%	79.77%
classifier 2	77.08%	86.89%
Averaging	77.19%	87.61%
MSE	75.95%	87.89%
CE	75.84%	87.89%
CFM	76.74%	90.67%
$CFM_{\delta_{min}}$	75.62%	90.11%

Table 3: Identification rates of different classification systems on SET2.

Classifiers	Test data	Validation data
classifier 1	79.82%	84.61%
classifier 2	85.32%	87.33%
Averaging	81.64%	89.50%
MSE	82.69%	89.72%
CE	82.98%	89.83%
CFM	88.07%	92.61%
$CFM_{\delta_{min}}$	86.14%	92.00%

Table 4: Identification rates of different classification systems on SET3.

Classifiers	Test data	Validation data
classifier 1	70.51%	79.50%
classifier 2	78.86%	84.00%
Averaging	81.08%	86.67%
MSE	80.51%	86.61%
CE	82.33%	86.61%
CFM	82.10%	90.28%
$CFM_{\delta_{min}}$	81.70%	89.83%

Table 5: Identification rates of different classification systems on SET4.

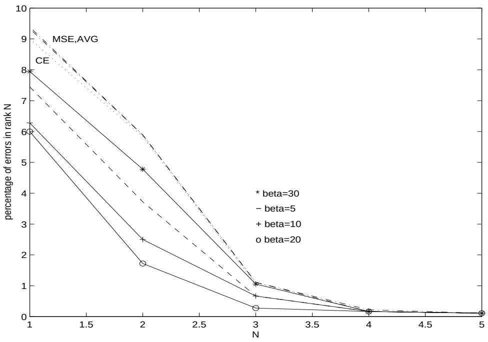

Figure 2: Ranking performance of CFM and some other combination schemes. The validation sessions are used for testing and the experiments are done on SET1.

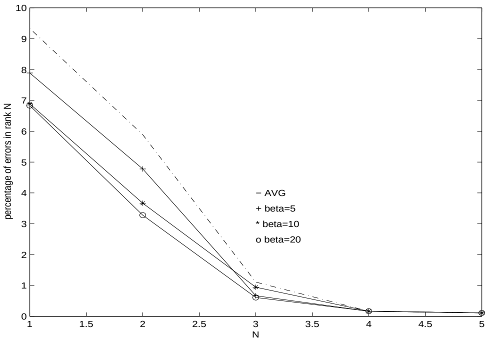

Figure 3: Ranking performance of $CFM_{\delta_{min}}$ and averaging schemes. The validation sessions are used for testing and the experiments are done on SET1.

ror, the correct speaker lies in most likely three speakers for all objective functions. The performance of the CFM objective function is better than the other objective functions when first most likely three speakers are considered. The performance of all OCC schemes are nearly identical if first 4 speakers are considered. The ranking performance of $CFM_{\delta_{min}}$ is given in Figure 3. The $CFM_{\delta_{min}}$ performance is given for three different values of β and the result of the averaging approach is also given for comparison. The performance of $CFM_{\delta_{min}}$ and averaging become identical when the percentage that the correct speaker lies in top 4 speakers is considered.

Experimental results have shown that CFM with $\beta = 20$ provided better performance compared to other OFs. We believe that this is not surprising since the most important point in classification system design is the maximization of the correct output and the values of the outputs of other incorrect speakers are not important from classification point of view, as far as they are less than the output of the correct speaker.

In [2], Hampshire *et al* stated that multiple neural networks trained with different objective functions may provide complementary behaviour. That is, their errors may not necessarily overlap. From the classifier combination point of view, this may provide a new way of developing MCSs. In other word, we may consider the results from different objective functions simultaneously in giving the joint decision. As a further research, the complementary behaviour of different objective functions should be analyzed.

7. REFERENCES

[1] S. Hashem. Optimal linear combinations of neural networks. *Neural Networks*, 10(4):599–614, 1997.

[2] J. B. Hampshire and A. H. Waibel. A novel objective function for improved phoneme recognition using time-delay neural networks. *IEEE Transactions on Neural Networks*, 1(2):216–228, June 1990.

[3] K. Al-Ghoneim and B.V.K. V. Kumar. Unified decision combination framework. *Pattern Recognition*, 31(12):2077–2089, 1998.

[4] N. Ueda. Optimal linear combination of neural networks for improving classification performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:207–215, 2000.

[5] H. Melin and J. Lindeberg. Guidelines for experiments on the POLYCOST database, January 1997.

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

[7] D. A. Reynolds and R. C. Rose. Robust text-independent speaker recognition using Gaussian mixture speaker nodels. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.

[8] H. Altınçay and M. Demirekler. An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification. *Speech Communication*, 30(4):255–272, April 2000.