

DECISION LEVEL FUSION OF USABLE SPEECH MEASURES USING CONSENSUS THEORY

Jashmin K. Shah, Brett Y. Smolenski, and Robert E. Yantorno
Temple University/ECE Dept. 1947 N. 12th St., Philadelphia, PA 19122-6077 USA
Email: shah@temple.edu, bsmolens@temple.edu, robert.yantorno@temple.edu
http://www.temple.edu/speech_lab

Abstract: The accuracy of speaker identification systems degrades when operating in an adverse acoustical environment due to a reduction in the quality of the speech being input into the system. Speech which is corrupted by stationary or non-stationary noise or interferer speech could be improved by taking only the portions which are minimally degraded but still useful for speaker identification systems. The extracted speech, which is also known as usable speech, could then be input into a speaker identification system to improve its accuracy. Several usable speech measures have been developed to identify usable segments from degraded speech. Unfortunately, the currently available measures only detect about 74% of usable speech, with 26% false alarms. To improve the classification rate, a decision level fusion technique, consensus theory, with an intelligent weighing scheme based on classifier's performance has been used. The experiment shows an improvement of 8% in correct detection and 22% reduction in false alarms from the best performing usable speech measure.

Key words: Usable Speech, Fusion System, Consensus Theory, Linear Opinion Pool, Logarithmic Opinion Pool, Intelligent Weighing Scheme.

1. Introduction

Information fusion is a very important task in pattern recognition as it is difficult to develop classifiers with a high identification performance rate. Different feature sets present different information and pattern class, and may therefore have different probabilistic models about the pattern class. In general, the aim of a classifier combination scheme is to take advantage of the strength of the individual classifiers and avoid their weakness, thereby improving the classification accuracy.

The technique of speaker identification involves the authentication of identification of a speaker from a large set of possible speakers. A speaker identification system must decide which speaker among an ensemble of speakers produced a given speech utterance. The performance of speaker identification system degrades under co-channel condition, i.e., when two people are talking at the same time. Speech signal of interest is termed as *target speech* and interfere is termed as *interferer speech*.

Previous studies have revealed that when the overall energy of the interfering speaker is the same as that of the target speaker, i.e. target-to-interferer ratio (TIR) is 0 dB, there exists 40% accuracy of a speaker identification system as compared with 90% correct identification in the case of no interferer [1]. This result seems reasonable because the interferer speech is not spread over the entire utterance, but only at certain portions of the co-channel speech. This result suggests that there are certain portions of speech which are minimally corrupted, but still useful for speech processing applications such as speaker identification. Further studies of TIR with speaker identification systems have determined that 20dB TIR is a reasonable lower limit for speaker identification to work reliably with co-channel as an input [1]. Usable speech measures determine the usability of co-channel speech in an operational environment since *a priori* information of TIR is not available.

Several measures have been developed to identify usable segments from degraded speech [2,3,4,5,6,7]. However, it is difficult to develop a classifier with a high identification performance rate due to the fact that the speech is diverse, and therefore, one is confronted with many different, energies, transition regions, and sounds which may not have steady state regions. To improve the performance of usable speech detection, it is necessary to fuse together a number of usable speech measures that provide complimentary information, in order to obtain reliable decisions.

The usable speech measures used in this research were Adjacent Pitch Period Comparison (APPC) [3] and Spectral Autocorrelation of Peak to Valley Ratio of LPC Residual (SAPVR Residual) [4].

1.1 Usable Speech Measure

The APPC measure takes advantage of the periodic structure of voiced single-speaker speech in the time domain. When there is little or no interference ($|TIR| > 20$ dB), the adjacent pitch periods of voiced speech are similar in 'shape', and therefore, the Euclidean distance between the adjacent pitch periods is small and the corresponding speech frame is declared usable. In the case of corrupted speech ($|TIR| < 20$ dB), there will be lack of periodicity due to the interferer speech, which will cause adjacent pitch periods to have different shapes. This results in a large Euclidean distance

between adjacent pitch periods, and the corresponding speech frame is then declared unusable.

SAPVR - Residual, which is a modified version of SAPVR [2], detects usable speech by examining the structure of the autocorrelation of the FFT of the LPC residual of the co-channel speech. When there is little or no interference, the autocorrelation of FFT produces structured harmonics with several peaks and valleys. If the ratio of the sum of the peaks to the first valley is above a given threshold, the corresponding speech frame is considered usable. In the case of corrupted speech, there will be a definite loss of structure in the autocorrelation of the FFT, which will produce fewer peaks and the ratio of the sum of the peaks to the first valley will fall below the given threshold, and therefore the frame will be considered unusable.

The threshold of measures has been chosen such a way that results in an equal amount of misses and false alarms. Correct detection of usable speech occurs when measure satisfy its threshold criteria and its $|TIR| > 20$ dB. False alarm occurs when measure declare as usable speech while $|TIR| < 20$ dB. It has been assumed that both measures will have complementary information since their probability density distribution are radically different (data now shown), and therefore, fusion of both measures should improve the usable speech detection.

Fusion systems are broadly classified as data level, feature level and decision level fusion [8]. Data level fusion is the most accurate fusion level among these three. However, it requires a high amount of computation at data level. In our case, data level fusion is not an option since a 40ms frame of 8 kHz sampled speech contains 320 sample points. Fusion of such a high dimensional amount of data would be computationally expensive and may not be possible in many practical situations. Feature level fusion occurs before the decision is taken. Recently two feature level fusion systems were developed. One using independent component analysis and non linear estimation [9], and the other is using Bayesian classification [10]. We develop decision level fusion which occurs after the decision taken by measure about specific pattern class. The proposed block diagram of a decision level fusion system is shown in Figure 1 below.

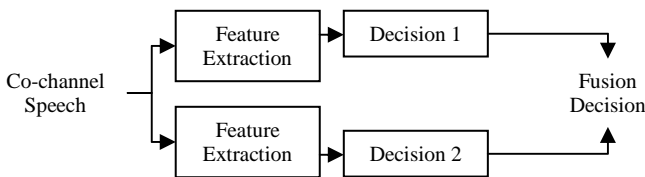


Figure 1: Proposed Block Diagram of Decision Level Fusion System.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background of consensus theory fusion approach and development of an intelligent weighing scheme. The experimental condition is given in Section 3 and results are presented in Section 4. Section 5 is summary and discussion. Section 6 outlines future area of research.

2. Background

Given two or more usable speech measures, the question is, how does one use the measures to provide an effective and reliable measure? To help with this problem, we have looked at the field of consensus theory, where the objective is to find a consensus amongst a group of experts (in our case measures). The classical approaches of linear opinion and logarithmic opinion pool consensus theory are discussed below.

2.1 Linear Opinion Pool Theory

One of the most frequently used consensus rule is the Linear Opinion Pool (Lin-Op), which is a linear weighted sum of the *a posteriori* probabilities of the classifiers [11]. For a given set of M classifiers (APPC and SAPVR Residual) and N pattern classes (Usable and Unusable), the general form of the Consensus Function (CF) for Lin-Op is defined as,

$$CF(p_1, \dots, p_N) = \sum_{m=1}^M w_m P(X = i | D_m = j) \quad (1)$$

where,

X = pattern data (TIR)

w_m = weighing factor

$P(X = i | D_m = j)$ represents the *a posteriori* probability density function that the tested data belongs to pattern class i when the decision of the m th classifier belongs to pattern class j .

2.2 Logarithmic Opinion Pool Theory

The combination scheme referred to as the Logarithmic Opinion Pool (Log-Op) described by [11] was derived using Bayes' theorem. The CF of Log-Op is defined as,

$$CF(p_1, \dots, p_N) = \sum_{m=1}^M \log w_m P(X = i | D_m = j) \quad (2)$$

The above equation (2) is the sum of weighted logarithmic *a posteriori* probability of classifiers. The modified form of the above equation can be expressed as,

$$CF(p_1, \dots, p_N) = \prod_{m=1}^M P(X = i | D_m = j)^{w_m} \quad (3)$$

The decision of equation (2) and equation (3) are expected to be same because of the logarithmic property.

Note that in the CF of Lin-Op and Log-Op above, the magnitude of the weight determines the influence of each classifier on the joint decision. The main problem in this combination scheme is the selection of the weights assigned to the classifiers. There are many heuristic approaches in the literature [11], however they ignore the reliability of the individual classifiers. Weighing operators can be classified in two ways, context dependent operators and context independent operators [12]. We have investigated the weighing scheme based on context dependent operators, which considers the classifier's class dependent and global reliabilities. Both reliabilities can be obtained using the information transmission theory which is discussed in next section.

2.3 Correspondence Between Detection Theory and Information Theory

The relation between a classification system and an information transmission system is similar to the one described in [14]. Observed data can be considered as an input to the transmission system and the output is the decision taken by the classifiers. The block diagram of classification system is shown in Figure 2.

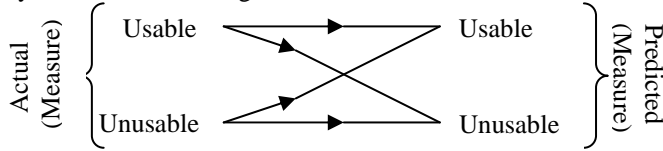


Figure 2: Information Transmission Channel Model.

The confusion matrix, using the above information transmission model to represent classification of each classifier, can be obtained, and is shown in Section 4. Note that the rows of the confusion matrix denote the values measured by the TIR, and the columns represent the values predicted by the measure. Using the elements of the confusion matrix, we can obtain the *a posteriori* probability $P(X = i | D_m = j)$. In the next section, we present the classifiers weight estimation theory, where the elements from confusion matrix and *a posteriori* probability are used.

2.4 Classifier Weights

In order to weigh the classifier's decision, we consider two propositions. One is that the classifier may be reliable when giving decision about one class and not reliable for another class. The other is that classifiers may have different global reliability. The classifier weights are estimated in an intelligent way in order to consider both class dependent and global classifier reliabilities simultaneously. The general form of the *m*th classifier's weight is defined as,

$$w_m = \alpha_m(j) \gamma_m \quad (4)$$

where,

$\alpha_m(j)$ represents class dependent classifier's reliability and γ_m represents global classifier's reliability.

The class dependent classifier reliability can be obtained by removing uncertainty of class from the distributed decided class. The available uncertainty can be obtain using Shannon's entropy equation,

$$H(X | D_m = j) = - \sum_{i=1}^N P(X = i | D_m = j) \log_2 P(X = i | D_m = j) \quad (5)$$

where,

$N = \text{pattern class} = 2$ (usable and unusable) and *a posteriori* probability $P(X = i | D_m = j)$ can be obtained using the confusion matrix.

By using equation (5), the class dependent classifier's reliability can be defined as,

$$\alpha_m(j) = \log_2 N - H(X | D_m = j) \quad (6)$$

The global classifier's reliability can be obtained by removing the average uncertainty about the channel input remaining after the channel output has been observed. This is obtained using the average residual uncertainty in the classifier, and can be defined as,

$$H(X | D_m) = - \sum_{i=1}^N \sum_{j=1}^N P(X = i, D_m = j) \log_2 P(X = i | D_m = j) \quad (7)$$

By using equation (7) above, global classifiers reliability can be obtained as,

$$\gamma_m = \log_2 N - H(X | D_m) \quad (8)$$

3. Experimental Condition

Experiments were performed on the decision of APPC and SAPVR Residual. The speech data used was taken from TIMIT database. To perform experiments on a large amount of data, 42 speech files, with equal numbers of male and female speaker were combined at 0dB TIR to obtain co-channel speech. All possible combination of 42 speech files, for a total of 861 co-channel sentences were obtained. Training session was performed on 430 co-channel files and tested using the remaining 431 co-channel files. Once the co-channel utterance was obtained, it was broken down into 40 ms frames with no overlap, since it has been

shown that usable speech characteristics have little dependence on overlap [1]. For each frame, the values of the measure, TIR, and spectral flatness were recorded. A decision of frames with greater than -35dB spectral flatness values was removed to exclude the unvoiced frames.

4. Experimental Results

To obtain the *a posteriori* probability for each classifier, the linear least squares fit of PDF data was obtained for usable and unusable distribution of SAPVR Residual and APPC on training session of 430 files, and the results are shown in Figure 3. Note that the data is plotted on a semilog scale to obtain the exponential equation.

Exponential equations for the *a posteriori* probability of usable and unusable data of the SAPVR Residual and APPC are obtained as;

$$\begin{aligned} P_{\text{SAPVR Residual}}(\text{usable}) &= 25.7e^{-0.6K} \\ P_{\text{SAPVR Residual}}(\text{unusable}) &= 8.6e^{-0.5K} \end{aligned} \quad (9)$$

$$\begin{aligned} P_{\text{APPC}}(\text{usable}) &= 0.26e^{-0.3K} \\ P_{\text{APPC}}(\text{unusable}) &= 0.05k^{0.16} \end{aligned} \quad (10)$$

where k is the respective measure value.

To perform the linear least squares fit of data for SAPVR - Residual, a measure values of 9 and above was chosen for usable and unusable since most correct detection of usable speech occurred in that region. For APPC, a measure value of 6 and below was chosen since most correct detection occurred in that region.

Note that, because the distribution of unusable of APPC was not exponential in nature, a log-log plot of linear least squares fit of data was performed to obtain the power equation shown in equation (10).

The confusion matrix for two classes, usable and unusable for 430 training files is,

$$CM_{\text{APPC}} = \begin{bmatrix} 0.74 & 0.26 \\ 0.41 & 0.59 \end{bmatrix} \quad (11)$$

$$CM_{\text{SAPVR Residual}} = \begin{bmatrix} 0.71 & 0.29 \\ 0.43 & 0.56 \end{bmatrix} \quad (12)$$

Note that the APPC has slightly better classification rate than

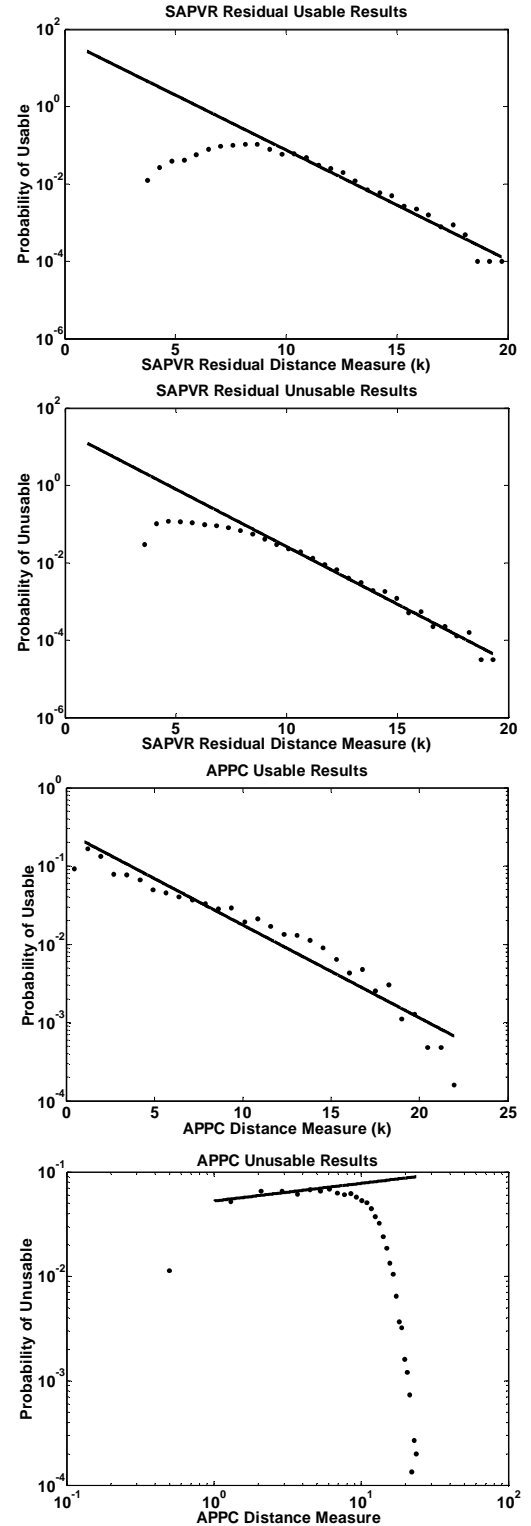


Figure 3: Linear Least Squares Fit of Data on Training Session: Upper two panels for SAPVR Residual usable and unusable respectively and lower two panels for APCC usable and unusable respectively.

SAPVR Residual. Using the above confusion matrix for each classifier and equation (7), classifier's class dependent reliability obtained as,

$$\alpha_{APPC}(usable) = 0.05, \alpha_{APPC}(unusable) = 0.08$$

$$\alpha_{SAPVR\ Residual}(usable) = 0.03, \alpha_{SAPVR\ Residual}(unusable) = 0.06$$
(13)

The global classifiers reliability γ_m was determined using equation (8) and are -

$$\gamma_{APPC} = 0.07, \gamma_{SAPVR\ Residual} = 0.05$$
(14)

The normalized final weight was obtained using equation (4) and are -

$$w_{APPC}(usable) = 0.62, w_{APPC}(unusable) = 0.60$$

$$w_{SAPVR\ Residual}(usable) = 0.38, w_{SAPVR\ Residual}(unusable) = 0.39$$
(15)

Using above final weight, *a posteriori* probabilities of the classifier shown in Figure 3 were weighted and combined using equation (2) and (3), and a new probability density distribution was obtained for linear and logarithmic opinion pool, and are shown in Figure 4 and 5 respectively.

Thresholds of 0.1 and 0.005 were selected for linear and logarithmic opinion pools respectively. Due to the application of the above thresholds, the probability of having hits has become higher than the probability of false alarms. The result of linear opinion pool and logarithmic opinion pool combination scheme is shown in Figure 6.

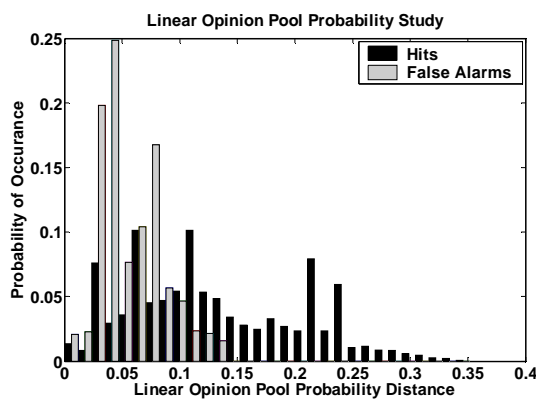


Figure 4: Probability Density Distribution for Linear Opinion Pool: Correct detections are shown in black and false alarms are shown in gray.

Using linear opinion pool, 80% correct detection of usable speech was obtained with 20% of false alarms. This account for 8% increase in correct detection with 26% reduction in the false alarms with compare to performance of APPC. The comparison of result of linear opinion pool with APPC is

shown in Figure 7. Note that the results of logarithmic opinion pool are not very encouraging. This is due the fact that the fused decision is a product (common) decision of both classifiers.

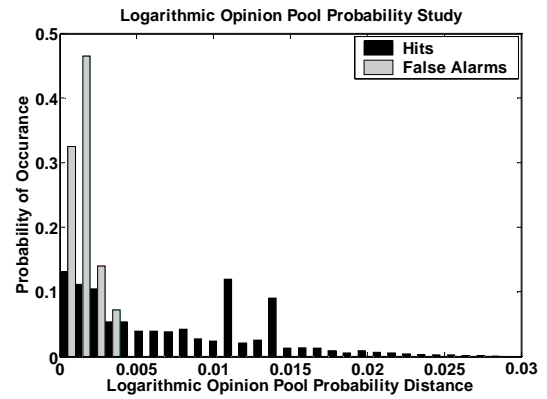


Figure 5: Probability Density Distribution for Logarithmic Opinion Pool: Correct detections are shown in black and false alarms are shown in gray.

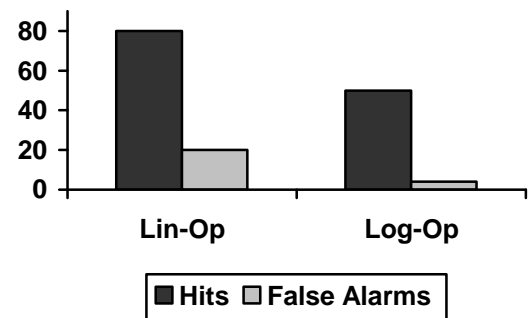


Figure 6: Comparison of Results of Fusion Techniques: Black bars show hit detection and gray bars shows false alarms.

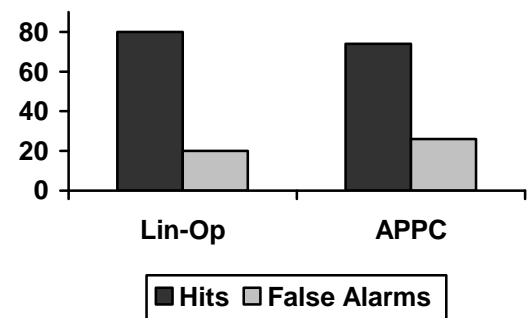


Figure 7: Comparison of Results of Linear Opinion Pool and APPC: Black bars show hit detection and gray bars shows false alarms.

5. Summary and Discussion

The problem of combining classifiers which uses different probabilistic representation of patterns to be classified was studied. It was noticed that fusion technique based on context dependent provides better classification compare to the heuristics weighing approach. The identification of usable speech was improved using two usable speech measures and a linear opinion pool combination scheme. If one had additional usable speech measures available to fuse, additional performance gains could be realized assuming that they provide complementary information.

6. Future Area of Research

Further improvement can be done by using more usable speech measures and a different weighing approach. One approach could be the use of soft composition, where different feature sets can always be simultaneously used in an optimal way to determine linear combination weights. In contrast to the concept where the more reliable measure takes the final decision, soft composition is a concept where all the rivals can work on the same task together, but the more reliable measure plays a more important role than others.

Acknowledgement

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-02-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Government purpose notwithstanding any copyright annotation thereon.

Disclaimer

The views and conclusions contained herein are those of the authors and should not be interpreted as necessary the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.

7. References

- [1]. Lovekin, J., Yantorno, R. E., Benincasa, S., Wenndt, S., and Huggins, M., "Developing Usable Speech Criteria for Speaker Identification". ICASSP, pp: 421-424, May 2001.
- [2]. Krishnamachari, K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J., "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp: 710-713, Nov. 2000.
- [3]. Lovekin, J., Krishnamachari, K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J., "Adjacent Pitch Period Comparison as a Usability Measure of Speech Segments Under Co-channel Conditions". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp: 139-142, Nov. 2001.
- [4]. Chandra, N., and Yantorno, R. E., "Usable Speech Detection Using the Modified Spectral Autocorrelation Peak to Valley Ratio Using the LPC Residual". 4th IASTED International Conference on Signal and Image Processing, pp: 146-149, Aug 2002.
- [5]. Kizhanatham, A. R., Yantorno, R. E., and Smolenski, B. Y., "Peak Difference Autocorrelation of Wavelet Transform Algorithm Based Usable Speech Measure". 7th World Multi-conference on Systemic, Cybernetics and Informatics, Aug 2003 (Submitted).
- [6]. Iyer, A. N., Gleiter, M., Smolenski, B. Y., and Yantorno, R. E., "Structural Usable Speech Measures Using LPC Residual". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, Dec 2003 (Accepted).
- [7]. Sundaram, N., Yantorno, R. E., Smolenski, B. Y., and Iyer, A. N., "Usable Speech Detection Using Linear Predictive Analysis - A Model Based Approach". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, Dec 2003 (Accepted).
- [8]. Hall, D. L., "Mathematical Technique in Multisensor Data Fusion". Artech House 1992.
- [9]. Smolenski, B. Y., Yantorno, R. E., and Wenndt, S. J., "Fusion of Co-channel Speech Measures Using Independent Components and Nonlinear Estimation". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, Nov. 2002.
- [10]. Smolenski, B. Y., Yantorno, R. E., "Fusion of Usable Speech Measures Using Quadratic Discriminant Analysis". IEEE International Symposium on Intelligent Signal Processing and Communication Systems 2003 (Submitted).
- [11]. Benediktsson, J. A., Swain, P. H., "Consensus Theoretic Classification Methods". IEEE Transaction on Systems, Man, and Cybernetics, Vol. 22, No. 4, Aug 1992.
- [12]. Altincay, H., and Demirekler, M., "An Information Theoretic Framework for Weight Estimation in the Combination of Probabilistic Classifiers for Speaker Identification". Speech Communication 30 (2000), pp: 255-272.
- [13]. Bloch, I., "Information Combination Operators for Data Fusion: A Comparative Review with Classification". IEEE Transaction on System, Man, and Cybernetics - Part A: Systems and Humans, Vol. 26, No. 1, Jan 1996.
- [14]. Hoballah, I. Y., and Varshney, P. K., "An Information Theoretic Approach to the Distributed Detection Problem". IEEE Transaction on Information Theory, Vol.35, No.5, Sep 1989.