Speaker Verification Using Orthogonal GMM with Fusion of Threshold, Identification Front-end, and UBM

Ningping Fan, Justinian Rosca, and Radu Balan

Real-time Vision and Modeling, Siemens Corporate Research Inc.
755 College Road East, Princeton, NJ, USA
{NingPing.Fan,Justinian.Rosca,Radu.Balan}@scr.siemens.com

Abstract

This paper shows that the performance of a Gaussian Mixture Model using a Universal Background Model (GMM-UBM) speaker verification (SV) system can be further improved by combining it with threshold and speaker identification (SI) "front-ends." The paper formulates performance in terms of false rejection rate and false acceptance rate of the overall SV system. We show analytically that an SI-based front-end can significantly decrease the false acceptance rate and only results in a slight increase in the false rejection rate. Experimentally we use a subset of NIST 2001 speaker recognition corpus with 10 registered speakers of 20 utterances against 10 imposters of 960 utterances. The results show significant reduction in the false acceptance rate, from 3.5% down to 1.0% (i.e. 71% error reduction) while maintaining the same zero false rejection rate.

1. Introduction

The state of the art GMM-UBM speaker verification system can be further improved. One way is to use discriminative kernels like the score-space kernel, which has reported to achieve 34% error reduction [1]. This paper focuses on another method to combine different verification techniques like the threshold and the identification front-end, and we have achieved 71% error reduction in experimentation.

Three speaker verification methods are shown in figure 1: (a) the threshold-based method, where the decision of rejection or acceptance is based on a threshold associated for each speaker GMM model which was chosen so that the false rejection rate (FR) and the false acceptance rate (FA) satisfy certain criterion, such as FR equals FA, or FA < 90%, etc. (b) the UBM-based method, in which a universal background model (UBM) [2] is created from all the registered speakers, and is used to against the individual models for binary classification. (c) the SI-based method, where the similarity score is calculated exhaustively for each speaker model in the database. The best matched model in agreement with the claimed identity can lead to accept decision. Here we used similarity to generalize the matching score, which can be the logarithm-likelihood in the GMM model [3] or the inverse of the distance measure in the VQ model [4].

This paper studies several approaches to combine these three different methods into a more robust decision system. We will show performance improvement in detail when the method (c) is used as a first stage process cascaded before the method (a) or (b). For that reason, method (c) is called the *identification front-end*. Then the method (a) and (b) are further employed either alone or combined together as follows to provide the best performance. If both (a) and (b) accept results in a final accept decision, and if either (a) or (b) reject results in a final reject decision. The same mechanism can also be used to enhance the speaker identification system, in which the verification threshold or an universal background model can be used to eliminate false identifications for test voices from unregistered speakers that have no corresponding voice models in the system database.

This paper is organized as follows. The section 2 first formulates the performance indexes of proposed system, and then the section 3 analyzes those equations for the speaker verification task. The section 4 presents the results from NIST 2001 speaker recognition corpus experiments. The section 5 draws conclusions.

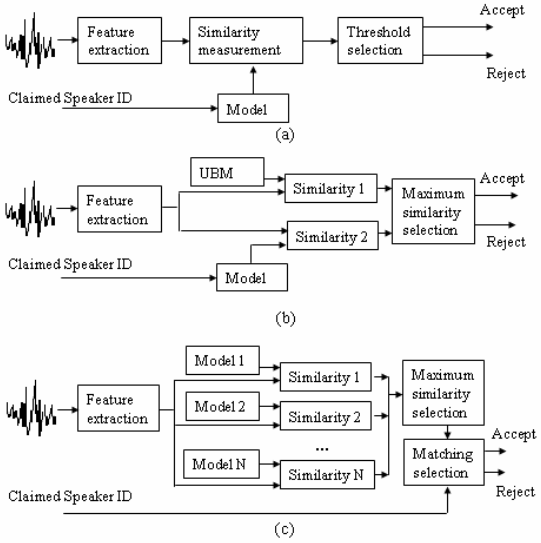

Figure 1 – Illustrations of speaker recognition systems; (a) the individual threshold-based speaker verification system; (b) the UBM-based speaker verification system; (c) the speaker identification (SI)-based speaker verification system.

2. Performance Formulation

For the simplicity, the following analysis assumes the combination of methods (c) and (a), but same analysis can also be applied to the combination of methods (c) and (b) if the log-likelihood scores were replaced with the log-likelihood ratios.

Let a speaker recognition system have N_M registered speaker models. Given N_I testing voices from registered speakers, and N_O testing voices from imposters, there can be $N_M * (N_I + N_O)$ possible matching scores $s(M)$ between a testing voice and a model M, as illustrated in figure 2.

Let N_{CA} be the total number of correct acceptances among N_I registered speakers' testing voices. Let N_{CR} be the total number of correct rejections among N_O imposters' testing voices. Let N_{CI} be the total number of correct identifications

among N_I registered speakers' testing voices. The performance indexes of a combined SV system of (c) and (a): the false rejection rate R_{FR}, the false acceptance rate R_{FA}, and the speaker identification rate R_{ID} can be calculated as follows.

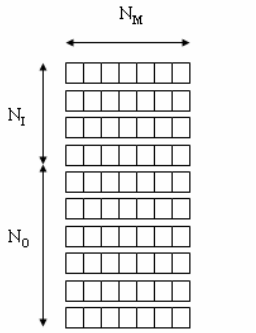

Figure 2 - All possible matching scores of a speaker recognition system for N_M voice models, N_I testing voices from registered speakers and N_O testing voices from imposters.

$$R_{FR} = 1 - \frac{N_{CA}}{N_I} \qquad N_I \to \infty \qquad \textbf{(1)}$$

$$R_{FA} = 1 - \frac{N_{CR}}{N_O N_M + N_I (N_M - 1)} \qquad N_I + N_O \to \infty \qquad \textbf{(2)}$$

$$R_{ID} = \frac{N_{CI}}{N_I} \qquad N_I \to \infty \qquad \textbf{(3)}$$

where the denominators are the total number of events that will happen and numerators are the true events. For example, the voice from imposters should all be rejected, thus the total number of events is $N_O N_M$. The voice from registered users should only be rejected for incorrect models, thus the total number of events is $N_I (N_M - 1)$, as in (2).

For threshold-based speaker verification system, the score is only calculated between a testing voice and a claimed identity voice model, and then compared with a threshold for acceptance or rejection. Its false rejection rate $R_{FR,T}(T)$ and false acceptance rate $R_{FA,T}(T)$ for a give threshold T are determined by the corresponding conditional probabilities as follows:

$$\begin{aligned} R_{FR,T}(T) &= 1 - \frac{N_{CA}}{N_I} \\ &= 1 - P(s(M) \ge T \mid v \in M) \qquad N_I \to \infty \end{aligned} \qquad \textbf{(4)}$$

$$\begin{aligned} R_{FA,T}(T) &= 1 - \frac{N_{CR}}{N_O N_M + N_I (N_M - 1)} \\ &= 1 - P(s(M) < T \mid v \notin M) \quad N_I + N_O \to \infty \end{aligned} \qquad \textbf{(5)}$$

For a combined speaker verification system that uses the identification front-end, the scores of a testing voice for all the voice models are calculated. If only the identification output or the model with the maximum score, agrees with the claimed identity model, the score is then compared with a threshold for further rejection of imposters who do not have a model in the system. Its false rejection rate $R_{FR,I}(T)$ and false acceptance rate $R_{FA,I}(T)$ for a threshold T are given as follows:

$$\begin{aligned} R_{FR,I}(T) &= 1 - \frac{N_{CI}}{N_I}\frac{N_{CA}}{N_I} \\ &= 1 - R_{ID} \cdot P(s(M) \ge T \mid v \in M) \qquad N_I \to \infty \end{aligned} \qquad \textbf{(6)}$$

$$\begin{aligned} R_{FA,I}(T) = 1 &- \frac{N_{CI}(N_M - 1)}{N_O N_M + N_I (N_M - 1)} \\ &- \frac{(N_I - N_{CI})(N_M - 2)}{N_O N_M + N_I (N_M - 1)} \\ &- \frac{N_O (N_M - 1)}{N_O N_M + N_I (N_M - 1)} \\ &- \frac{N_O P(\hat{s} < T \mid v \notin M)}{N_O N_M + N_I (N_M - 1)} \qquad N_I \to \infty \end{aligned} \qquad \textbf{(7)}$$

$$\textit{where} \quad \hat{s} = \max s(M) \; \forall M$$

where (6) is due to that only correctly identified voice from registered users using the identification-based front-end can be further correctly accepted using the threshold-based SV system. The equation (7) can be explained as follows. The first term is due to the fact that the correct identification of a voice from registered users will also reject it from the rest $N_M - 1$ wrong models, and the total number of events of correct identification of voices from registered users is $N_{CI}(N_M - 1)$. The second term is due to the fact that even an incorrect identification of a voice from registered users will also correctly reject it from the rest $N_M - 2$ wrong models, and the total number of events of incorrect identification of voices from registered users is $(N_I - N_{CI})(N_M - 2)$. The third term is due to the fact that the identification process will always correctly reject a voice from imposters for $N_M - 1$ wrong models, and the total number of these events is $N_O(N_M - 1)$. The fourth term is due to the fact that the threshold-based verification as in (5) is only for the imposters and thus the total number of events is the total number of imposters times the probability of correct rejection.

3. Performance Analysis

To compare the two verification systems, we can compute and compare the performance indexes (4) versus (6) and (5) versus (7). One can derive that (7) is a monotonic decreasing function with respect to N_M, as the partial derivative of $R_{FA,I}(T)$ with respect to N_M is always less than zero, and furthermore the $R_{FA,I}(T)$ will approach zero as N_M approaches infinity.

Assuming a uniform posterior distribution of the identification results:

$$\frac{\partial R_{FA,I}(T)}{\partial N_M} = -\frac{(N_{CI}+N_I)N_I(1-R_{ID})}{\left(N_O N_M + N_I(N_M-1)\right)^2} - \frac{(N_{CI}+N_I)N_{CI}(1-P(\hat{s}<T \mid v\notin M))}{\left(N_O N_M + N_I(N_M-1)\right)^2} < 0 \quad (8)$$

$$R_{FA,I}(T) = 1 - \frac{N_{CI}N_M + (N_I - N_{CI})N_M + N_O N_M}{N_O N_M + N_I N_M} = 0 \qquad N_M \to \infty \quad (9)$$

Therefore, it can be concluded that the following is true in general.

$$R_{FR,I}(T) \geq R_{FR,T}(T)$$
$$R_{FA,I}(T) << R_{FA,T}(T) \qquad when \quad N_M >> 1 \quad (10)$$

In other words, the identification front-end can drastically decrease the false acceptance rate but with a small increase of false rejection rate. The latter can be compensated up to the identification rate limit, i.e. $1-R_{ID}$, by lowering T so that $P(s(M)>T \mid v\in M)\to 1$. If $P_0 = P(s(M)\geq T \mid v\in M)$, $P_1 = P(s(M)<T \mid v\notin M)$, $P_2 = P(\hat{s}(M)<T \mid v\notin M)$, the figure 3 top plot assumes constant probabilities, while in the bottom one, P_2 is decreasing proportionally with N_M, and constant others. In any case, (10) is indeed true.

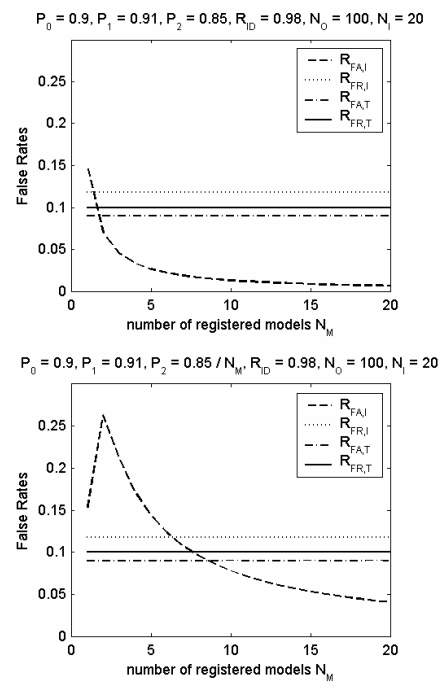

Figure 3 – Theoretic speaker verification performance indexes plots: (a) P_0=0.9, P_1=0.91, P_2=0.85, R_{ID}=0.98, N_O=100, N_I=20; (b) P_0=0.9, P_1=0.91, P_2=0.85/N_M, R_{ID}=0.98, N_O=100, N_I=20.

4. Experiments

Speech signals from 20 different identities (10 male and 10 female) are taken from NIST 2001 speaker recognition corpus in "devtest-training" section. 10 of them are used as registered users, and 10 are used as imposters. The wave files are cut into several pieces each 10 ~ 20 sec length, 16 bit and 8 KHz. For the 10 registered users, 2 pieces are used for testing, 4 ~ 8 pieces are used for orthogonal GMM [5] individual model training, and rest 2 ~ 3 pieces are grouped together for UBM training. For 10 imposters, all the pieces are used for testing, and each piece will test all the 10 registered users.

All the speech signals are converted into feature vectors of MFCC coefficients [6]. 10 individual models and one UBM are created from the training feature vectors, corresponding to registered speakers, with id listed in the first row of table 1. Each testing utterance from both registered speakers and imposters is then used to compare with each model to generate a normalized logarithm likelihood score, as shown in table 1.

Table 1 – Normalized log-likelihood scores of a GMM system with 10 registered models with id listed in the first row and 980 testing voices of 20 from registered users and 960 from imposters. Threshold values are underlined.

id	4	5	6	9	10	12	14	15	16	17	UBM
1	-1.19	-1.2	-1.31	-1.25	-1.2	-1.41	-1.52	-1.57	-1.46	-1.52	-1.18
1	-1.08	-1.12	-1.15	-1.08	-1.05	-1.2	-1.36	-1.39	-1.28	-1.32	-1.03
1	-1.24	-1.25	-1.31	-1.32	-1.25	-1.49	-1.59	-1.69	-1.52	-1.59	-1.24
1	-1.27	-1.34	-1.35	-1.36	-1.32	-1.37	-1.46	-1.57	-1.44	-1.46	-1.24
1	-1.25	-1.27	-1.37	-1.31	-1.26	-1.43	-1.55	-1.63	-1.48	-1.53	-1.22
1	-1.15	-1.19	-1.2	-1.19	-1.16	-1.29	-1.45	-1.46	-1.36	-1.43	-1.12
1	-1.11	-1.15	-1.21	-1.16	-1.14	-1.31	-1.44	-1.46	-1.36	-1.42	-1.11
1	-1.15	-1.16	-1.22	-1.21	-1.15	-1.3	-1.43	-1.49	-1.36	-1.46	-1.11
1	-1.09	-1.13	-1.13	-1.15	-1.1	-1.21	-1.38	-1.38	-1.31	-1.37	-1.07
1	-1.13	-1.19	-1.13	-1.19	-1.1	-1.16	-1.3	-1.34	-1.26	-1.3	-1.07
1	-1.22	-1.2	-1.29	-1.26	-1.22	-1.41	-1.53	-1.58	-1.45	-1.54	-1.19
2	-1.57	-1.62	-1.55	-1.77	-1.72	-1.48	-1.54	-1.57	-1.43	-1.56	-1.42
2	-1.57	-1.6	-1.56	-1.8	-1.7	-1.74	-1.63	-1.86	-1.55	-1.6	-1.38
2	-1.59	-1.65	-1.59	-1.81	-1.76	-1.51	-1.6	-1.62	-1.46	-1.55	-1.41
2	-1.49	-1.51	-1.53	-1.72	-1.65	-1.43	-1.45	-1.54	-1.38	-1.44	-1.31
2	-1.3	-1.25	-1.33	-1.43	-1.32	-1.38	-1.3	-1.5	-1.33	-1.31	-1.22
2	-1.47	-1.41	-1.47	-1.66	-1.48	-1.61	-1.62	-1.82	-1.53	-1.67	-1.38
2	-1.48	-1.47	-1.47	-1.69	-1.56	-1.49	-1.54	-1.68	-1.45	-1.55	-1.35
2	-1.26	-1.33	-1.3	-1.3	-1.36	-1.28	-1.24	-1.41	-1.29	-1.23	-1.17
2	-1.3	-1.34	-1.34	-1.45	-1.37	-1.43	-1.34	-1.6	-1.35	-1.37	-1.24
2	-1.55	-1.51	-1.55	-1.69	-1.57	-1.69	-1.7	-1.9	-1.6	-1.72	-1.5
2	-1.63	-1.6	-1.62	-1.8	-1.66	-1.84	-1.79	-2.01	-1.69	-1.81	-1.56
3	-1.18	-1.23	-1.25	-1.26	-1.22	-1.24	-1.27	-1.42	-1.26	-1.24	-1.15
3	-1.17	-1.21	-1.25	-1.24	-1.22	-1.24	-1.25	-1.45	-1.24	-1.22	-1.13
3	-1.18	-1.16	-1.36	-1.38	-1.25	-1.33	-1.33	-1.53	-1.31	-1.33	-1.15
3	-1.24	-1.26	-1.42	-1.46	-1.32	-1.4	-1.39	-1.57	-1.39	-1.4	-1.24
3	-1.28	-1.26	-1.4	-1.47	-1.34	-1.42	-1.45	-1.62	-1.42	-1.43	-1.27
3	-1.41	-1.39	-1.47	-1.49	-1.42	-1.49	-1.5	-1.66	-1.49	-1.46	-1.38
3	-1.13	-1.17	-1.16	-1.18	-1.15	-1.13	-1.19	-1.37	-1.16	-1.16	-1.07
3	-1.32	-1.28	-1.46	-1.49	-1.32	-1.49	-1.48	-1.65	-1.47	-1.5	-1.29
3	-1.2	-1.21	-1.38	-1.38	-1.23	-1.38	-1.38	-1.58	-1.38	-1.4	-1.18
3	-1.17	-1.2	-1.36	-1.39	-1.25	-1.37	-1.37	-1.57	-1.37	-1.35	-1.17
4	-1.06	-1.26	-1.26	-1.29	-1.24	-1.24	-1.3	-1.4	-1.31	-1.3	-1.13
4	-1.02	-1.15	-1.23	-1.22	-1.14	-1.26	-1.32	-1.44	-1.29	-1.34	-1.09
5	-1.38	-1.23	-1.55	-1.51	-1.39	-1.64	-1.67	-1.78	-1.61	-1.64	-1.35
5	-1.25	-1.19	-1.3	-1.38	-1.32	-1.39	-1.48	-1.47	-1.41	-1.48	-1.2
6	-1.26	-1.3	-1.09	-1.25	-1.21	-1.37	-1.45	-1.48	-1.42	-1.45	-1.12
6	-1.18	-1.2	-1.02	-1.15	-1.1	-1.34	-1.45	-1.45	-1.39	-1.43	-1.08

7	-1.14	-1.14	-1.25	-1.25	-1.18	-1.31	-1.29	-1.4	-1.3	-1.3	-1.15
7	-1.1	-1.09	-1.22	-1.21	-1.14	-1.25	-1.25	-1.33	-1.26	-1.26	-1.11
7	-1.13	-1.16	-1.3	-1.28	-1.2	-1.33	-1.32	-1.44	-1.33	-1.34	-1.17
7	-1.12	-1.15	-1.29	-1.26	-1.17	-1.34	-1.36	-1.45	-1.35	-1.34	-1.14
7	-1.16	-1.2	-1.33	-1.31	-1.23	-1.39	-1.42	-1.47	-1.4	-1.4	-1.2
7	-1.17	-1.18	-1.26	-1.29	-1.22	-1.3	-1.33	-1.38	-1.32	-1.31	-1.17
7	-1.06	-1.06	-1.16	-1.17	-1.11	-1.18	-1.13	-1.25	-1.17	-1.16	-1.06
8	-1.4	-1.59	-1.5	-1.62	-1.71	-1.27	-1.29	-1.33	-1.28	-1.23	-1.21
8	-1.47	-1.63	-1.58	-1.7	-1.77	-1.31	-1.38	-1.38	-1.33	-1.31	-1.29
8	-1.45	-1.64	-1.59	-1.73	-1.78	-1.31	-1.34	-1.34	-1.3	-1.29	-1.26
8	-1.27	-1.43	-1.38	-1.45	-1.58	-1.22	-1.19	-1.33	-1.19	-1.14	-1.14
8	-1.33	-1.52	-1.49	-1.51	-1.65	-1.3	-1.21	-1.44	-1.27	-1.16	-1.16
8	-1.45	-1.6	-1.6	-1.74	-1.76	-1.37	-1.34	-1.45	-1.32	-1.3	-1.27
8	-1.35	-1.54	-1.47	-1.57	-1.7	-1.26	-1.24	-1.3	-1.25	-1.2	-1.2
9	-1.26	-1.24	-1.13	-0.99	-1.15	-1.36	-1.47	-1.51	-1.44	-1.48	-1.08
9	-1.16	-1.11	-1.03	-0.65	-0.96	-1.16	-1.33	-1.34	-1.24	-1.34	-0.86
10	-1.32	-1.31	-1.39	-1.39	-1.11	-1.54	-1.65	-1.78	-1.56	-1.6	-1.2
10	-1.35	-1.31	-1.44	-1.39	-1.16	-1.58	-1.67	-1.77	-1.59	-1.61	-1.24
11	-1.3	-1.21	-1.48	-1.49	-1.28	-1.53	-1.58	-1.7	-1.54	-1.54	-1.32
11	-1.26	-1.21	-1.45	-1.48	-1.24	-1.48	-1.5	-1.66	-1.47	-1.48	-1.27
11	-1.33	-1.26	-1.5	-1.53	-1.34	-1.58	-1.58	-1.74	-1.54	-1.58	-1.34
11	-1.26	-1.18	-1.46	-1.47	-1.26	-1.53	-1.53	-1.69	-1.49	-1.51	-1.28
11	-1.35	-1.26	-1.5	-1.58	-1.32	-1.59	-1.56	-1.74	-1.55	-1.57	-1.34
11	-1.35	-1.26	-1.49	-1.57	-1.34	-1.57	-1.52	-1.74	-1.53	-1.57	-1.33
12	-1.48	-1.63	-1.5	-1.63	-1.68	-1.29	-1.47	-1.54	-1.46	-1.39	-1.3
12	-1.22	-1.35	-1.27	-1.35	-1.32	-0.98	-1.22	-1.34	-1.23	-1.2	-0.98
13	-1.66	-1.74	-1.68	-1.89	-1.87	-1.6	-1.55	-1.58	-1.51	-1.51	-1.42
13	-1.57	-1.72	-1.66	-1.73	-1.85	-1.49	-1.42	-1.49	-1.49	-1.38	-1.35
13	-1.43	-1.51	-1.48	-1.57	-1.61	-1.4	-1.32	-1.38	-1.38	-1.34	-1.26
13	-1.49	-1.53	-1.54	-1.68	-1.63	-1.47	-1.4	-1.41	-1.41	-1.43	-1.31
13	-1.57	-1.65	-1.64	-1.77	-1.78	-1.49	-1.44	-1.47	-1.45	-1.43	-1.34
14	-1.45	-1.49	-1.49	-1.7	-1.59	-1.4	-1.22	-1.49	-1.37	-1.38	-1.27
14	-1.39	-1.49	-1.42	-1.59	-1.59	-1.29	-1.18	-1.4	-1.33	-1.28	-1.2
15	-1.43	-1.52	-1.44	-1.59	-1.58	-1.35	-1.34	-1.26	-1.33	-1.34	-1.29
15	-1.39	-1.51	-1.41	-1.53	-1.58	-1.31	-1.32	-1.22	-1.32	-1.31	-1.24
16	-1.32	-1.36	-1.36	-1.59	-1.52	-1.19	-1.17	-1.16	-1	-1.2	-1.05
16	-1.44	-1.5	-1.48	-1.69	-1.63	-1.34	-1.31	-1.3	-1.11	-1.31	-1.16
17	-1.6	-1.8	-1.67	-1.84	-1.87	-1.54	-1.5	-1.59	-1.49	-1.36	-1.36
17	-1.74	-1.95	-1.85	-1.98	-2.02	-1.77	-1.61	-1.89	-1.69	-1.43	-1.47
18	-1.33	-1.4	-1.3	-1.43	-1.49	-1.17	-1.17	-1.12	-1.2	-1.11	-1.08
18	-1.47	-1.59	-1.51	-1.66	-1.69	-1.42	-1.37	-1.37	-1.36	-1.29	-1.25
18	-1.38	-1.41	-1.35	-1.57	-1.52	-1.27	-1.29	-1.22	-1.23	-1.31	-1.2
18	-1.31	-1.33	-1.25	-1.46	-1.38	-1.22	-1.26	-1.18	-1.2	-1.24	-1.14
18	-1.28	-1.27	-1.24	-1.43	-1.37	-1.24	-1.25	-1.19	-1.2	-1.23	-1.11
18	-1.47	-1.56	-1.44	-1.6	-1.63	-1.34	-1.33	-1.34	-1.36	-1.23	-1.21
18	-1.41	-1.46	-1.36	-1.56	-1.56	-1.26	-1.28	-1.2	-1.26	-1.23	-1.19
19	-1.43	-1.67	-1.43	-1.55	-1.73	-1.22	-1.32	-1.38	-1.35	-1.23	-1.14
19	-1.43	-1.66	-1.43	-1.56	-1.68	-1.2	-1.33	-1.39	-1.4	-1.24	-1.17
19	-1.58	-1.9	-1.65	-1.76	-1.95	-1.59	-1.45	-1.79	-1.6	-1.33	-1.27
19	-1.4	-1.61	-1.41	-1.53	-1.61	-1.2	-1.32	-1.42	-1.34	-1.25	-1.14
19	-1.68	-1.96	-1.75	-1.8	-2.04	-1.59	-1.48	-1.84	-1.64	-1.37	-1.35
19	-1.28	-1.49	-1.26	-1.4	-1.51	-1.07	-1.23	-1.24	-1.22	-1.16	-1.06
19	-1.32	-1.52	-1.3	-1.45	-1.53	-1.07	-1.27	-1.24	-1.25	-1.21	-1.1
19	-1.32	-1.54	-1.36	-1.46	-1.52	-1.17	-1.29	-1.43	-1.33	-1.22	-1.13
20	-1.35	-1.38	-1.39	-1.56	-1.47	-1.25	-1.24	-1.21	-1.19	-1.21	-1.13
20	-1.46	-1.5	-1.49	-1.71	-1.66	-1.36	-1.33	-1.32	-1.26	-1.34	-1.21
20	-1.47	-1.46	-1.48	-1.67	-1.59	-1.36	-1.36	-1.33	-1.26	-1.4	-1.26
20	-1.46	-1.51	-1.48	-1.69	-1.63	-1.34	-1.31	-1.3	-1.26	-1.3	-1.18
20	-1.43	-1.47	-1.47	-1.69	-1.59	-1.29	-1.27	-1.29	-1.24	-1.25	-1.16
20	-1.45	-1.45	-1.43	-1.64	-1.58	-1.35	-1.38	-1.34	-1.25	-1.43	-1.23

The minimum score of each testing voice compared with its own model is selected as the threshold for a testing voice to be accepted or rejected by that identity voice model, as underlined in the table 1. Thus all the testing voice from registered users will be correctly accepted for each voice model. The corresponding verification results for the threshold-based, UBM based, and various combinations with the identification-based front-ends speaker verification systems are shown in table 2.

Table 2 - Results of various algorithms for the previous data

Method	FR	FA
Individual Threshold:	1-20/20 = 0	1-808/960 = 0.158
Identification+Threshold	1-20/20 = 0	1-836/960 = 0.129
UBM	1-20/20 = 0	1-926/960 = 0.035
Identification+UBM	1-20/20 = 0	1-939/960 = 0.022
Identification+Threshold+UBM	1-20/20 = 0	1-950/960 = 0.010

5. Conclusions

An approach combining the power of three different speaker verification methods has been proposed. We show analytically that the combined solution is better than the standalone solution. Experimentation using a subset of standard NIST speaker recognition corpus also provides strong evidence to support that conclusion.

Because the proposed method is based on decision fusion, the same analysis can also be applied to the more advanced score-space kernel based systems. However we expect only a moderate enhancement because the complimentary characteristics will less effective when the classifier approaching its limit.

6. References

[1] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machine", IEEE trans on Speech and Audio Processing, 13-2, pp 203-210, Mar. 2005.

[2] A. D. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, No 1, pp. 19-41, Jan. 2000.

[3] A. D. Reynolds and C. R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, 3(1): pp. 72-83, 1995.

[4] N. Fan and J. Rosca; "Enhanced VQ-based algorithms for speech independent speaker identification", Proc. AVBPA03, pp. 470-477, Guildford, UK, Jun. 2003.

[5] L. Liu and J. He; "On the use of orthogonal GMM in speaker recognition", Proc. ICASSP99, pp. 845-848, Phoenix, USA, Mar. 1999.

[6] S. Kim, Yonsei, T. Eriksson, H-G Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition" Proc. ICASSP04, pp. 405-408, Montreal, Canada, May 2004.

[7] W. Lim and N. S. Kim, "Bayesian approach to text-independent speaker verification", International Conference on Speech Processing, Aug. 2001.

[8] S. Z. Li, D. Zhang, C. Ma, H. Y. Shum, and E. Chang, "Learning to boost GMM based speaker verification", Proc. Eurospeech03, pp. 1677-1680, Geneva, Switzerland, 2003.