VOL. 12, NO. 2, JANUARY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CROSS MATCH-CHMM FUSION FOR SPEAKER ADAPTATION OF VOICE BIOMETRIC

A. K. Ariff, Sh-Hussain Salleh, Kamarulafizam I. and Alias Mohd. Noor
Centre for Biomedical Engineering, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
E-Mail: armouris@gmail.com

ABSTRACT

The most significant factor affecting automatic voice biometric performance is the variation in the signal characteristics, due to speaker-based variability, conversation-based variability and technology variability. These variations give great challenge in accurately modeling and verifying a speaker. To solve this variability effects, the cross match (CM) technique is proposed to provide a speaker model that can adapt to variability over periods of time. Using limited amount of enrollment utterances, a client barcode is generated and can be updated by cross matching the client barcode with new data. Furthermore, CM adds the dimension of multimodality at the fusion-level when the similarity score from CM can be fused with the score from the default speaker modeling. The scores need to be normalized before the fusion takes place. By fusing the CM with continuous Hidden Markov Model (CHMM), the new adapted model gave significant improvement in identification and verification task, where the equal error rate (EER) decreased from 6.51% to 1.23% in speaker identification and from 5.87% to 1.04% in speaker verification. EER also decreased over time (across five sessions) when the CM is applied. The best combination of normalization and fusion technique methods is piecewise-linear method and weighted sum.

Keywords: cross match, speaker recognition, speaker adaptation, speaker verification, voice biometric.

INTRODUCTION

Most of the time, voice displays both behavioral and physiological characteristics (Bolle *et al.*, 2003). It is well understood that besides physical traits such as the shape of vocal tract and vocal cords, the voice also depends on the behavioral features of the speaker, such as the state of mind. Furthermore, no one will repeat the same words in exactly the same way. That is why speech signals have various variability - i) Speaker-based variability, ii) Conversation-based or language variability and iii) technology variability. These variations make the automatic speaker authentication with high accuracy a difficult task to be achieved. The system has to decide the source of variability, whether it is due to intra-speaker (same speaker) or inter-speaker (different speakers).

One of the solutions to the problems mentioned above is to apply speaker adaptation technique during the training process of the voice biometric to 'learn and record' all the variability in the speech samples. In order to maintain high accuracy as long as possible, the reference model has to be adapted or updated with variability by applying adaptation techniques. Adaptation techniques have also been shown to enhance the performance of voice biometric system, especially when systems are provided with only limited enrolment data or to convert speaker independent models to speaker dependent models. This is because over a period of time, the speech signal can vary significantly, as well as differ systematically in some respect from the training speech.

Traditional goal of speaker adaptation is to convert a general speaker-independent model into a specific speaker-dependent model, by using the limited enrolment data (Hazen *et al.*, 2003). The adaptation process normally a very fast process as a priori knowledge of the parameters to be adapted is used, depending on the amount of training data available. Maximum likelihood linear regression (Leggetter and Woodland, 1995) and maximum a posteriori (Gauvian and Lee, 1994) are the most popular adaptation techniques. There are many variants of these techniques can be found in many literature, such as in (Mari'ethoz and Bengio, 2002) and (Ahn *et al.*, 2002), for text-independent and text-independent speaker verification system respectively.

CROSS MATCH

A new type of adaptation technique called cross match (CM) technique was proposed by Sheikh Hussain in 1997 for speaker verification system for determining the intra-speaker variation based on similarity match of two speech signals. In this technique, the speech frames are grouped into corresponding samples by evaluating the local similarities within the speech frames based on correlation score. By averaging, a client barcode (CB) is created by merging these samples into a sequence of vectors. The cross match technique is also combined with vector quantization called VQ-CM by Alwi in 2004. The new technique gave good results with the improvement of 64.85% using speaker specific threshold. A. K. Ariff in 2008 made a different set-up in the combination of CM, where the combination is combined based on the multi-modal biometric method of score fusion of discrete HMM and CM.

One of the major weaknesses in the current CM technique is the fact that procedure is based on the merging of only two samples at a time. The combined two samples will then be merged with the next samples based on the correlation score, until a single CB is created. This final CB is actually the combination of two sample pairing, rather than the collective grouping of all samples. In this work, new approaches are proposed to improve the CM, particularly in the stage of CB model creation.

SCORE FUSION

The CM and CHMM are combined based on the multimodal biometrics concept of score-fusion. Multimodal biometrics or combination of two or more biometrics is introduced to counter the weaknesses of using unimodal biometric (Ratha *et al.*, 2001). There are normally four levels of fusion - score, feature, decision and sample level (Kumar *et al.*, 2003). In score fusion method, individual scores are fused into a single score. Due to its simplicity, matching-score fusion is the most popular fusion method, besides giving a very good performance and intuitiveness (Indovina *et al.*, 2003). In many cases, when matchers' scores from vendor systems are available, the fused score from more than one system can be used to evaluate the performance of the multimodal biometric system in the same manner as a single biometric system.

Kittler *et al.* (1998) evaluated several classifier combination rules on voice, frontal face and face profile biometrics. The sum of a posteriori probabilities rule outclassed other fusion rules due to its resilience to errors. In evaluation of a system combining fingerprint, face and hand geometry, Jain *et al.* (2005) investigate various combinations of fusion methods - simple sum, min-score and max-score with normalized scores (min-max, median-MAD, tanh, double sigmoid, Parzen and z-score). They concluded that (a) when densities are unknown, tanh is better than min-max and z-score; (b) simple sum fusion of z-score, tanh and min-max normalization outperformed other combinations; and (c) customized user-by-user weighting is better than generic weightings. Different set of experiments are also conducted by Snelick *et al.* (2005), combining min-max, tanh, z-score and adaptive normalization methods with max score, min score, simple sum and user weighting fusion. The results show that the best combination is min-max normalization and simple-sum fusion.

There is no general conclusion on the best combination of normalization method and fusion technique. Based on the data available and the application, experiments have to be carried out to find the answer. The combination that works well in an environment and modalities is not necessarily the same case when it is applied in other system with different environment and conditions. In this work, based on two classifiers or matchers, the best combination of normalization technique and fusion method for voice biometric system will be concluded at the end of experiments.

RESEARCH METHODOLOGY

The designed biometric system consists of two sessions; the enrollment session and the authentication or recognition session. In the enrollment session, a new speaker or client (with known identity) is enrolled or registered into the system. The speaker is prompted to utter the digits 'zero' to 'nine' in Malay language with the repetition of five times of each digit. Using fixed high-quality microphone in office environment, the database is collected in five sessions. Five tokens are recorded in each session for a period of six month.

Then, the features of the recorded speech are extracted using Mel-Frequency Cepstrum Coefficient (MFCC) to be used to train the Continuous Hidden Markov Model (CHMM). A model topology of 5-state continuous HMM (CHMM) is used to represent each digit model. At the same time, the Cross Match (CM) will build the client barcode from the same 12-vector MFCC features. Both models will be used to represent the clients and are stored in the database. The threshold for the client is set by evaluating the impostor speech utterances with the speaker utterances to determine the equal error rate (EER) threshold. It involves the CHMM and CM scores calculation, followed by the score normalization and score fusion. This threshold will be stored for use during recognition phase.

After the enrollment, the speakers are allowed to use the system in the recognition session, either in identification or verification stage. At the identification stage, the unknown user will be identified based on a 1 : N comparison, i.e. the utterance will be compared with the all clients' models stored in the database, while at the verification stage, the comparison is based on 1 : 1, i.e. the utterance will be compared with the identified client's model only. The decision threshold will be applied at both stages in order to accept or reject the speaker.

A. Multimodal score fusion of CHMM and CM

When a speech feature is cross matched with the client barcode, the cross match (CM) correlation score of the speech feature and the barcode will be calculated. This score will be combined with the continuous HMM (CHMM) score using various methods. This sort of score combination is one of multimodal biometric setup called score-level fusion. Both scores need to be normalized before the fusion takes place, as shown in Figure-1.

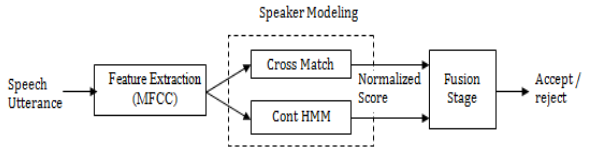

Figure-1. Score fusion of CHMM and CM.

Five common normalization methods will be implemented, namely min-max, z-score, tanh estimator, median and median absolute deviation (MAD) and piecewise-linear (Ribaric and Fratric, 2005), given in formulas 1 to 5 below. All normalization methods will be combined with five different fusion techniques - simple sum, weighted sum, product, minimum and maximum (refer to Table-1), giving 25 combinations, where the best combination will be determined.

$$s_k^{'} = \frac{s_k - \min}{\max - \min} \quad (1)$$

$$s_k' = \frac{s_k - \mu}{\sigma} \quad (2)$$

$$s_k' = \frac{s_k - median}{MAD} \quad (3)$$

$$s_k' = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{s_k - \mu}{\sigma}\right)\right)+1\right\} \quad (4)$$

$$s_k' = \begin{cases} 0 & \text{if } s_k < \min(O_i^G) \\ \dfrac{s_k - \min(O_i^G)}{\max(O_i^I) - \min(O_i^G)} & \text{if } \min(O_i^G) < s_k < \max(O_i^I) \\ 1 & \text{if } s_k > \max(O_i^I) \end{cases} \quad (5)$$

Table-1. Score fusion techniques.

Score fusion	Formula
Simple Sum	$\sum_{i=1}^{N} S_i$
Product Rule	$\prod_{i=1}^{N} S_i$
Minimum Score	Min (S_1, S_2, …., S_N)
Maximum Score	Max (S_1, S_2, …., S_N)
Weighted Sum	$W_1.S_1 + W_2 . S_2 + ….. + W_N . S_N$ where W is the weight

B. Cross match - Similarity matching

Cross match technique consists of two steps, which are similarity matching followed by barcode generation. In similarity matching step, two samples of the speech tokens are compared at all possible different relative positions. The similarity between the two samples of $X_{1j,}$ and X_{2j} when the relatives position of sample X_{2j} with respect to sample X_{1j} is p is given by the correlation score, $r(X_{ij}, X_{2j}, p)$. The two samples X_{1j} and X_{2j} are almost similar when the value $r(X_{ij},X_{2j},p)$ is high. Otherwise, if the value is negative, that indicates dissimilarity between the two samples. In other words, the best match between the two samples X_{1j} and X_{2j}.is indicated by the maximum value of $r(X_{ij},X_{2j},p)$ that occurs at the corresponding value of p. The correlation can be represented as:

$$r(X_1, X_2, p) = \frac{\sum (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{\sqrt{\sum (X_1 - \bar{X}_1)^2 \sum (X_2 - \bar{X}_2)^2}} \quad (6)$$

As an alternative to the simple correlation r value above, a different approach of using t value or student's t-distribution function is implemented, as the unique properties of both distributions prevent the occurrence of high correlation score at the start or the end of the similarity matching process. In this new matching correlation, the complex formula of t-distribution are simplified respectively in the form of

$$similarity = r(X_{1j}, X_{2j}, p)\sqrt{\frac{n_{ij} - 2}{1 - r(X_{1j}, X_{2j}, p)^2}} \quad (7)$$

where n is the number of frames of overlaps at this position.

Assume that the N tokens are arranged in such a way as to form a group such that the samples within the group match each other at relative positions that indicate similarity. Tokens X_{1j} and X_{2j} are cross matched to form a group in which token X_{1j}, and X_{2j} are most similar. At position $p(X_{1j}, X_{2j})$, where the token X_{1j}, and X_{2j} are most similar, both tokens will be cross matched to form a group of sequence. By grouping two tokens in this way at each time, the next step is to merge the two tokens within a group to form a single sequence, as illustrated in Figure-2. After having found a sequence representing the two tokens, a further attempt is made to form a group comprising N tokens for the client by the same process. The procedure is repeated several times until all the training tokens are used up to form a group representing the client barcode.

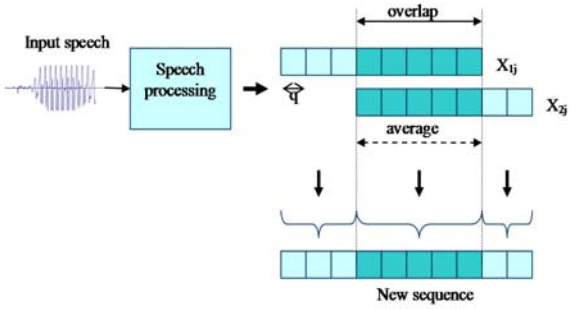

Figure-2. Averaging at the most similar position.

To generate client barcode from N tokens, the first token, X_1 is cross matching with second token, X_2 in order to find the relative position p, the highest correlation position. In the first cross matching, a combined sequence of X_1 and X_2 called *sequence 1* of feature vectors was obtained. In the next cross matching, the *sequence 1* will be cross matched with the third feature vectors to generate a new sequence called *sequence 2*. This process is repeated until the final sequence, *sequence N* is obtained to be taken as client barcode (CB), as shown in Figure-3.

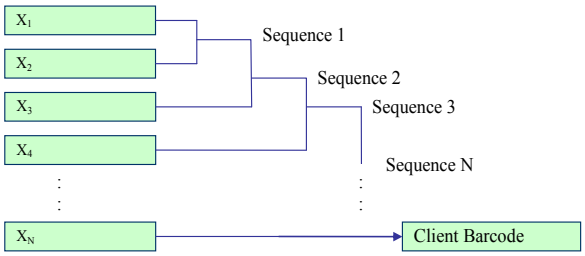

Figure-3. Client barcode generation.

In the above CM method, the similarity is calculated locally by merging two samples, and updated with other samples, one by one. Rather than merging two samples at a time, the new idea is proposed by merging all the samples globally, merging together all samples at one time. In this strategy, the speech sample with the longest frame length is used as the reference sample. All other samples will be compared with this reference sample only, and the relative position, *p* where the similarity score is the highest is recorded. After all samples are matched, based on the *p*, the overall samples group position will be determined, such as in the Figure-4.

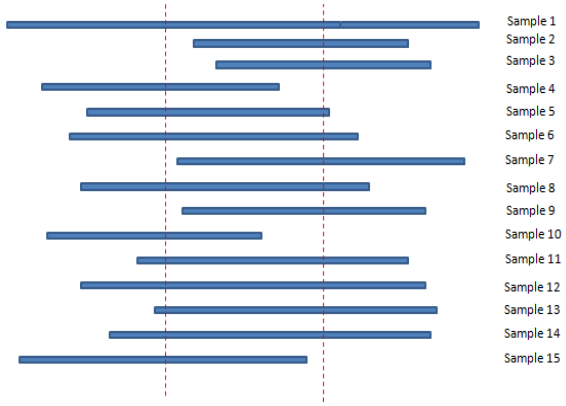

Figure-4. The longest frame (sample 1) as reference.

RESULTS AND ANALYSIS

In this section, the performances of the voice biometric system based on CHMM and CHMM-CM fusion are evaluated. The overall performance for both speaker identification and verification is measured based on the equal error rate (EER). The performances are compared between two system designs.

a) Evaluation on speaker identification

For every digit, a CHMM model was built representing all the clients. Table-2 shows that for identification results, the average EER is 6.51% ranging from 5% to 7.5%. Digit 7, 8 and 9 gave an overall best performance, while digit 2, 4 and 6 provide good performance result. However, digit 0, 1, 3 and 5 show the worst performance.

Table-2. CHMM identification performance.

Digit	EER (%)
0	7.29
1	7.54
2	6.13
3	7.61
4	6.24
5	7.33
6	6.17
7	5.98
8	5.65
9	5.18
Average	**6.51**

In CHMM-CM fusion, two scores from both CHMM and CM were normalized and fused together becoming a single score. The combination of five normalization methods and five fusion techniques were studied. All together there were 25 combinations. Table-3 shows the result for this score-fusion multimodal system.

It can be seen from Table-3 that all the fusion combinations gave better results compared to the previous experiment which is based on only CHMM. The average identification EER of CHMM is 6.51% but based on the average column of the table, all fusion EERs are less than this value. These findings confirm with results of other research that adding multimodality will make the biometric system more robust, reliable and accurate. The best combination is the piece-wise linear normalization and weighted sum fusion, which gave the lowest EER value of 1.23%.

b) Evaluation on speaker verification

The task of speaker verification is to verify the true client from impostor speakers. For each client, CHMM and CM model is built for every digit. The EER for each digit is determined and for all digits, all EERs are averaged to give the overall result. Based on the finding in identification results, the combination of piece-wise linear and weighted sum fusion is used, and the result is compared with the baseline EER results from CHMM-only experiment.

Table-4 and Figure-5 show the verification performance based on CHMM and CHMM-CM. In general, based on CHMM, the average EER score is 5.87%, ranging from 4.5% to 7%. But when CHMM is combined with CM, the results show a huge drop of 79% reduction of EER, from 5.87% EER to 1.23% EER. As it can be seen in from Figure 5, there is a big gap between both EER values.

c) Performance degradation over sessions

In this experiment, the identification performance of CHMM models and CHMM-CM fusion models are

evaluated in every session. The models were updated after every session as new speech utterances are enrolled.

Table-3. CHMM-CM fusion identification performance.

		0	1	2	3	4	5	6	7	8	9	Average	
	HMM	7.29	7.54	6.13	7.61	6.24	7.33	6.17	5.98	5.65	5.18	**6.51**	
Fusion	Normalize	EER (%)											
Simple sum	Min-max	1.83	1.95	1.58	1.87	1.68	1.87	1.59	1.98	1.85	1.56	1.78	**1.91**
	z-score	1.72	1.81	1.64	1.74	1.70	1.77	1.63	1.84	1.76	1.65	1.73	
	tanh	2.13	1.91	1.88	1.92	1.75	1.98	1.73	2.09	2.07	1.88	1.93	
	MAD	2.24	2.33	2.15	2.27	2.06	2.23	2.07	2.34	2.22	1.95	2.19	
	piecewise	1.99	1.95	1.86	2.06	1.88	1.99	1.81	2.11	1.97	1.74	1.94	
Weighted sum	Min-max	1.65	1.68	1.56	1.79	1.61	1.67	1.49	1.83	1.62	1.54	1.64	**1.50**
	z-score	1.78	1.83	1.67	1.73	1.70	1.86	1.71	1.96	1.79	1.65	1.77	
	tanh	1.47	1.51	1.32	1.53	1.35	1.38	1.29	1.43	1.39	1.21	1.39	
	MAD	1.58	1.53	1.35	1.54	1.44	1.61	1.47	1.54	1.42	1.35	1.48	
	piecewise	1.24	1.35	1.14	1.26	1.21	1.27	1.15	1.32	1.24	1.08	**1.23**	
Product	Min-max	2.01	2.18	1.80	2.04	1.83	2.13	1.83	1.95	1.88	1.79	1.94	**1.74**
	z-score	1.60	1.53	1.47	1.72	1.50	1.67	1.44	1.62	1.48	1.41	1.54	
	tanh	1.82	1.89	1.61	2.04	1.68	1.94	1.79	1.95	1.77	1.75	1.82	
	MAD	1.56	1.52	1.39	1.49	1.37	1.48	1.40	1.49	1.45	1.34	1.45	
	piecewise	2.08	2.04	1.95	2.03	1.82	2.00	1.91	2.05	1.93	1.73	1.95	
Minimum	Min-max	1.83	1.79	1.57	1.87	1.69	1.77	1.63	1.79	1.85	1.58	1.74	**1.96**
	z-score	1.37	1.43	1.15	1.36	1.21	1.38	1.15	1.29	1.32	1.19	1.29	
	tanh	1.77	1.83	1.66	1.86	1.55	1.79	1.64	1.72	1.71	1.67	1.72	
	MAD	2.13	2.07	1.78	2.09	1.74	2.09	1.80	2.11	1.88	1.86	1.96	
	piecewise	1.81	1.72	1.48	1.68	1.43	1.77	1.56	1.73	1.61	1.58	1.64	
Maximum	Min-max	1.77	1.83	1.44	1.74	1.57	1.71	1.55	1.68	1.63	1.59	1.65	**1.49**
	z-score	1.39	1.47	1.24	1.38	1.18	1.44	1.19	1.35	1.36	1.21	1.32	
	tanh	1.42	1.39	1.29	1.44	1.24	1.39	1.18	1.43	1.35	1.13	1.33	
	MAD	1.52	1.58	1.36	1.57	1.29	1.64	1.33	1.58	1.52	1.38	1.48	
	piecewise	1.79	1.69	1.49	1.70	1.59	1.77	1.55	1.81	1.88	1.59	1.69	

Table-4. Verification performance comparison.

Client	EER (%)	
	CHMM	CHMM - CM
1	6.42	1.27
2	5.32	1.33
3	5.36	0.87
4	4.89	0.93
5	5.03	0.89
6	4.92	1.04
7	4.55	0.83
8	5.23	0.93
9	6.78	1.03
10	5.91	0.98
11	6.44	1.08
12	6.74	1.39
13	7.13	1.45
14	6.41	0.79
15	6.93	1.84
Average	**5.87**	**1.04**

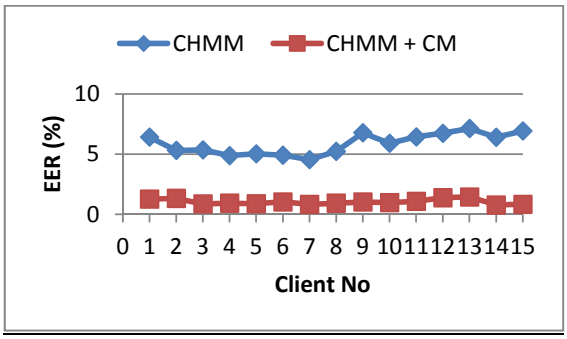

Figure-5. CHMM and CHMM-CM verification performance.

With more data gathered after every additional session, there will be increment in speech variability in the database.

Based on the results in Table-5 and Figure-6, the performance of CHMM worsens from session to session as the EERs are increased. This shows that CHMM alone is not adapted to the new added variability. But fusing the CHMM and CM gave a completely different results - over the sessions, the performance became better and better as the EER decreased. This is due to two reasons: the models are adapted with the all the variability of the speech features and the score-fusion implementation increases the accuracy. This proves the role of CM as a good adaptation technique and it also added multimodality effect to the voice biometric system.

Table-5. Average performance over sessions.

Session	EER (%)	
	CHMM	CHMM - CM
1	3.43	2.04
2	4.32	1.95
3	5.57	1.36
4	5.96	1.33
5	6.51	1.23

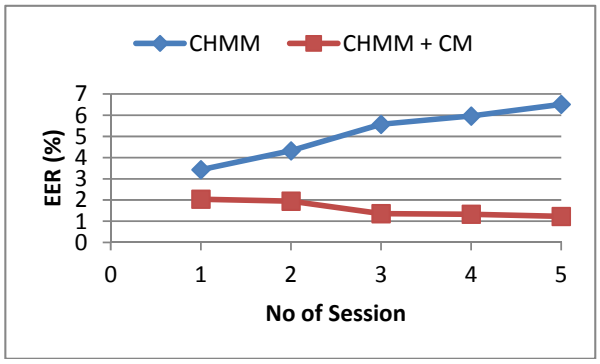

Figure-6. Identification performance over sessions.

CONCLUSIONS

It is highlighted in the beginning that the automatic voice biometric system can be affected significantly by the signal characteristics variation. Variability issue also occurs when voice samples which are recorded within a session are highly correlated compared with samples recorded in different sessions. Thus, the speaker model created solely from a single session (e.g. enrolment session) can experience problems when the system is used over extended periods of time. The last experiment shows this kind of property over sessions, but the proposed CM technique managed to adapt the variability and improved the performance.

We have demonstrated that by combining CM and CHMM by fusing both scores gave significant improvements in identification and verification results. Five normalization methods and five fusion techniques are tested, but the best combination for both system is piece-wise linear normalization and weighted-sum fusion.

Due to the fact that the CM is based on the similarity matching, the technique has potential to be applied in other applications involving biomedical signals such as ECG, EEG and heart sound where the signals will be accurately modeled if the time alignment variation is reduced.

ACKNOWLEDGEMENTS

This research is supported by CBE (Center for Biomedical Engineering) at Universiti Teknologi Malaysia, funded by Minister of Science, Technology and Innovation (MOSTI), Malaysia under Science Fund grant (R.J130000.7945.4S127).

REFERENCES

A. K. Ariff. 2008. Speaker Adaptation Based on Cross Match Technique, Universiti Teknologi Malaysia.

Ahn S., Kang S. and Ko H. 2000. Effective Speaker Adaptation for Speaker Verification. International Conference on Speech and Signal Processing. 1: 1081-1084.

Alwi M. 2004. Speaker Recognition Based on Cross Match Technique, Universiti Teknologi Malaysia.

Bolle R. M., J. H. Connell, S. Pankanti, N. K. Ratha and A. W. Senior. 2003. Guide to Biometrics, New-York: Springer-Verlag.

Gauvain J. L. and Lee C.H. 1994. Maximum Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains. IEEE Transaction on Speech and Audio Processing. 291-298.

Hazen, T. J., Jones, D. A., Park, A., Kukolich, L. C., and Reynolds, D. A. 2003. Integration of Speaker Recognition into Conversational Spoken Dialogue Systems, European Conference on Speech Communication and Technology. 1961-1964.

Indovina M., U. Uludag, R. Snelick, A. Mink and A. Jain. 2003. Multimodal Biometric Authentication Methods: A COTS Approach, Workshop on Multimodal User Authentication, Santa Barbara, CA.

Jain A. K., K. Nandakumar and A. Ross. 2005. Score Normalization in Multimodal Biometric Systems, Pattern Recognition. 38(12): 2270-2285.

Kittler, M. Hatef, R.P. Duin, J.G. Matas. 1998. On Combining Classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence. 226-239.

Kumar A., D. C. M. Wong, H. C. Shen and A. K. Jain. 2003. Personal Verification Using Palmprint and Hand Geometry Biometric, Proceeding of 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA). 668-678.

Leggetter, C. and Woodland, P. C. 1995. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM. Computer Speech and Language, 171-185.

Mari'ethoz J. and Bengio S. 2002. A Comparative Study for Adaptation Methods for Speaker Verification. International Conference on Spoken Language Processing, 581-584.

Ratha N. K., J. H. Connell, and R. M. Bolle. 2001. Enhancing Security and Privacy in Biometrics-based

Authentication Systems. IBM Systems Journal. 40(3): 614-634.

Ribaric S. and Fratric I. 2005. A Matching-score Normalization Technique for Multimodal Biometric Systems, Proceedings of Third COST 275 Workshop - Biometrics on the Internet. 55-58.

Sheikh Hussain Shaikh Salleh. 1997. An Evaluation of Preprocessors for Neural Network Verification, University of Edinburgh.

Snelick R., U. Uludag, A. Mink, M. Indovina and A. Jain. 2005. Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(3): 450-455.