EFFECTS OF TIME LAPSE ON SPEAKER RECOGNITION RESULTS

*Homayoon Beigi**

Recognition Technologies, Inc.
3616 Edgehill Road
Yorktown Heights, NY 10598, USA
beigi@RecognitionTechnologies.com

ABSTRACT

The effect of time lapse has not been studied well in most biometrics. Here, this effect is studied for Speaker Recognition, namely, Speaker Identification and Speaker Verification. The RecoMadeEasy^{TM} speaker recognition engine has been used to obtain baseline results for 22 speakers who have been involved in a long-term study. The speakers have given data in three seatings with 1 to 2 months delay between consecutive collections. The speakers were real proficiency test candidates who were asked to speak in response to prompts. At each seating, several recordings were made in response to different prompts. The error rates are discussed, going from one seating to the next, for Identification and Verification. Large degradations are seen across different seatings. Two different adaptation techniques have been studied for reducing this discrepancy with very promising results.

***Index Terms*—** speaker recognition, speaker identification, speaker verification, time lapse, speaker adaptation

1. INTRODUCTION

The effect of time lapse has not been studied well in biometrics. Although the literature is full of brief discussions about time lapse effects in speaker recognition, no proper quantitative study has been done on the subject.[1, 2] There are two main types of time lapse effects: *short-term* and *long-term* (aging). Here, short-term effects are studied for two functions of speaker recognition, namely speaker identification and speaker verification.[3]

Speaker identification starts with the modeling of the vocal characteristics of speakers based on their sample speech (called enrollment data) and storing them in a database. Then, given a new speech excerpt, the recognition engine returns the identity of the speaker from the database. If the engine allows for a no-match result, the process is called open-set identification, otherwise, it is a closed-set identification process. If at the time of testing, the identity of the speaker is presented to the engine along with the speech sample, then the sample is matched against the model for that speaker in the database and some other model(s) representing competing speakers. If the sample matches the speaker's model better than the competing model(s), then the speaker is verified, otherwise the speaker is rejected. This process is called verification. The RecoMadeEasy[1] speaker recognition engine has been used to obtain baseline results for 22 speakers who have been involved in a persistent (ongoing) study.[3]

Speakers were involved in language proficiency testing where they had to repeat their tests due to undesirable scores. The speakers used here retook their tests two more times after the original testing was accomplished. The time lapse between consecutive tests was on the average between 1 to 2 months. Table 1 provides the dates for the first, second and third tests for each candidate. The words test and trial will be used interchangeably from this point on. Each test consists of multiple audio segments which are each about 1 minute long. These segments are free-form responses to questions to assess the candidates' proficiency in the English language. Unfortunately, due to the fact that this type of study is quite rare, no standard corpus is available.

The RecoMadeEasy^{TM} Speaker Recognition engine uses a Gaussian Mixture Model (**GMM**) (see [3]) approach to conduct identification and verification of the speakers. Under normal circumstances, the first response of the first test (first trial) is used to enroll the speaker in the database. Consequent segments are identified or verified against the enrollment data captured from the first response. This scenario is used to conduct the rest of the test without any need for a proctor, therefore reducing the cost of testing. The data is obtained from a real-world application and has not been manipulated or specifically collected for this purpose. These are real candidates taking tests to be evaluated for English proficiency.

*Homayoon Beigi is the President of Recognition Technologies, Inc. and Adjunct Professor of Mechanical Engineering at Columbia University

[1]RecoMadeEasy^{TM} is the Commercial Speaker Recognition Engine of Recognition Technologies, Inc.

Speaker No.	First Trial	Second Trial	Third Trial
1	2007/08/12	2007/09/09	2007/11/25
2	2007/08/19	2007/10/14	2007/12/06
3	2007/08/05	2007/09/09	2007/10/14
4	2007/10/07	2007/11/11	2007/12/09
5	2007/08/19	2007/10/07	2007/12/02
6	2007/08/16	2007/10/11	2007/12/07
7	2007/08/23	2007/10/14	2007/12/09
8	2007/08/12	2007/09/09	2007/10/07
9	2007/08/12	2007/10/07	2007/12/02
10	2007/09/09	2007/10/10	2007/11/25
11	2007/08/16	2007/10/11	2007/12/06
12	2007/08/12	2007/10/12	2007/12/05
13	2007/08/16	2007/10/11	2007/12/06
14	2007/08/23	2007/10/14	2007/12/02
15	2007/08/12	2007/09/09	2007/10/07
16	2007/08/12	2007/10/07	2007/12/02
17	2007/08/12	2007/10/08	2007/12/07
18	2007/08/14	2007/10/07	2007/12/06
19	2007/08/14	2007/10/08	2007/12/07
20	2007/08/22	2007/10/14	2007/12/07
21	2007/08/12	2007/10/14	2007/12/02
22	2007/08/14	2007/10/12	2007/12/03

Table 1. Time of Audio Capture for Individual Speakers

Tests have revealed that there is substantial degradation in the results of both open-set identification and verification from one seating to the next. There are two main reasons for this degradation. The first well-known reason is known as channel mismatch between the enrollment session and the recognition of consequent tests. Many different approaches have been taken to reduce the effects of this mismatch.[3] Channel mismatch has in the past been mostly connected to handset mismatch, however, it is quite more complicated than that. In addition to handset mismatch, changes in the ambient noise (sometimes called source noise [4]), acoustic properties of the ambience (such as echo and reverberation), microphone distance and positioning (angle), strain on the vocal tracts (holding the handset on one's shoulder) and many more are also responsible for these types of mismatches.[5]

Different techniques have been proposed for handling this type of mismatch by considering specific sources of mismatch. These include Handset Score Normalization (H-Norm) [6], feature mapping [7], and speaker model synthesis (SMS) [8]. Others have approached the problem by suppressing the effects through using the T-Norm, the Z-Norm [9, 10] and Feature Warping techniques [11].

The second reason for degradation is a combination of other factors such as physiological changes, environmental changes, emotional changes, etc.[12, 13] These effects are not very well understood and are bundled here in one category called time lapse effects. Among these changes, there are some which get worse with time. We are interested in these effects which are collectively called *time lapse effects*. Note that we are not dealing with what the literature calls aging, since aging deals with much longer effects, outside the range of these shorter-term studies. Aging effects deal with more of the physiological changes that affect speakers as substantial time progresses.[12, 13]

To see these effects of interest we have considered three consecutive tests per candidate. The changes between the first test and the second test include both channel-mismatch and time lapse effects. However, by doing a third test and seeing further degradation of the recognition results, we can conclude that time lapse effects have caused most of the extra degradation seen from the first trial to the third trial as compared to the changes from the first trial to the second trial.

First, a description of the data is given in the following section. Then, we discuss these degradations in more detail by doing a quantitative analysis of the Identification and Verification results. Following this discussion, we try to reduce the effects of time lapse using several adaptation techniques and the results are reported for identification and verification tasks followed by concluding remarks.

2. THE AUDIO DATA

The audio data was collected using the μ-Law amplitude coding technique [14] at a sampling rate of 8 kilo Hertz (kHz). The audio was then immediately converted to the High-Efficiency Advanced Audio Coding Format (**HE-AAC**) [15] which is a very aggressive, lossy and low-bit-rate audio compression technique. **HE-AAC** was used to stream the audio to a server through flash. The audio, in turn, was converted back to μ-Law 8-kHz audio and subsequently converted to a 16-bit linear Pulse Code Modulation (**LPCM**) which was used in the recognizer for enrollment, identification and verification purposes.

3. A GAUSSIAN MIXTURE MODEL RECOGNIZER

The RecoMadeEasy^{TM} speaker recognition engine was used for obtaining results. This engine is a **GMM**-based text-independent and language-independent engine. It uses models for the speaker and the competing models to conduct the identification and verification tasks. The population in the

identification task is the 22 speakers described in the next section plus competing models. The models are parameters for collections of multi-variate normal density functions which describe the distribution of the Mel-Cepstral features [3] for speakers' enrollment data. This distribution is represented by Equation 1.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \quad (1)$$
$$where \begin{cases} \mathbf{x},\ \boldsymbol{\mu} \in \mathscr{R}^d \\ \mathbf{\Sigma} : \mathscr{R}^d \rightarrow \mathscr{R}^d \end{cases}$$

In 1, $\boldsymbol{\mu}$ is the mean vector where,

$$\boldsymbol{\mu} \stackrel{\Delta}{=} \mathscr{E}\{\mathbf{x}\} \stackrel{\Delta}{=} \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x}) d\mathbf{x} \quad (2)$$

The so-called "Sample Mean" approximation for 2 is,

$$\boldsymbol{\mu} \approx \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{x}_i \quad (3)$$

where N is the number of samples and $\mathbf{x}_i$ are the Mel-Cepstral feature vectors [3].

The Variance-Covariance matrix of a multi-dimensional random variable is defined as,

$$\mathbf{\Sigma} \stackrel{\Delta}{=} \mathscr{E}\left\{(\mathbf{x}-\mathscr{E}\{\mathbf{x}\})\,(\mathbf{x}-\mathscr{E}\{\mathbf{x}\})^T\right\} \quad (4)$$
$$= \mathscr{E}\left\{\mathbf{x}\mathbf{x}^T\right\} - \boldsymbol{\mu}\boldsymbol{\mu}^T \quad (5)$$

This matrix is called the Variance-Covariance since the diagonal elements are the variances of the individual dimensions of the multi-dimensional vector, $\mathbf{x}$. The off-diagonal elements are the covariances across the different dimensions. Some have called this matrix the Variance matrix. Mostly in the field of Pattern Recognition it has been referred to simply as the Covariance matrix which is the name we will adopt here.

The Unbiased estimate of $\mathbf{\Sigma}$, $\tilde{\mathbf{\Sigma}}$ is given by the following expression,

$$\tilde{\mathbf{\Sigma}} = \frac{1}{N-1}\sum_{i=0}^{N-1}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T \quad (6)$$
$$= \frac{1}{N-1}\left[\mathbf{S}_{xx} - N(\boldsymbol{\mu}\boldsymbol{\mu}^T)\right] \quad (7)$$

where the sample mean $\boldsymbol{\mu}$ is given by equation 3 and the second order sum matrix, $\mathbf{S}_{xx}$ is given by,

$$\mathbf{S}_{xx} = \sum_{i=0}^{N-1} \mathbf{x}_i\mathbf{x}_i^T \quad (8)$$

[3] describes details of a **GMM**-based recognizer.

4. BASELINE SPEAKER RECOGNITION

As previously mentioned, each candidate goes through a testing procedure in which questions are asked and responses from the candidate are recorded. Under usual circumstances, the first audio response is used to enroll the speaker into the system. All the responses average to about 1 minute of audio. Figure 1 shows the results of identification of individuals among the 22 candidates in our database. All subsequent audio responses are identified at a rate of 100% (an error rate of 0%). In this case, although the enrollment and recognition data differ, there is no channel mismatch. These results are expected from a good commercial recognition system. However, as conditions change and the candidates return to be tested for a second or third time, a substantial degradation is noted, see figure 1. For the second trial (test) and the third trial, there is channel mis-match [3, 6] present as well as time lapse effects. Since the channels are chosen at a completely random manner in both second and third trials, the extra degradation seen between trial 2 and trial 3 is most likely due to time lapse effects.

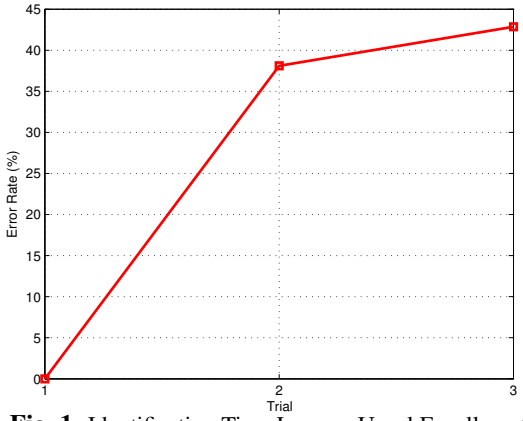

Fig. 1. Identification Time Lapse – Usual Enrollment

Figure 2 shows similar results for the verification process. In this figure, three Detection Error Tradeoff (**DET**) curves [16] are presented using the first response in the first test for enrollment, consequent data in the first test for verification of trial 1 and the second response in the second and third trials for verification of those trials. The plot shows results which are similar to those seen in the identification case. Namely,

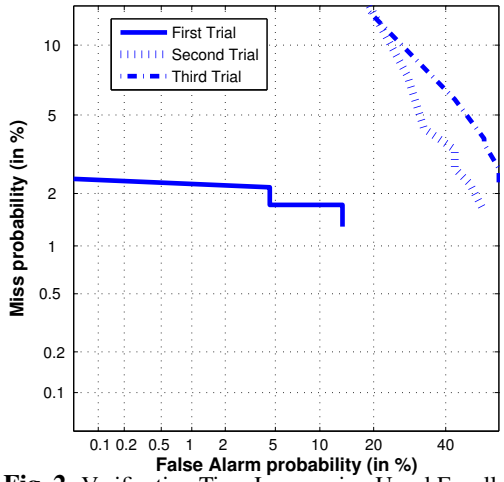

Fig. 2. Verification Time Lapse using Usual Enrollment

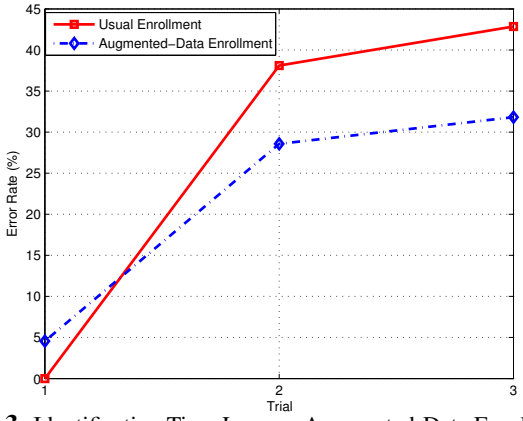

Fig. 3. Identification Time Lapse – Augmented-Data Enrollment

the Equal Error Rate (**EER**) increases from about 2.5% to nearly an order of magnitude higher for the second and third trials. The **EER** is defined as the point were the missed probability equals the false alarm probability in the **DET** curves. In consistence with the results for the identification tests, the performance of the verification system also degrades in time as we move from the second to the third trial, whereas, the two channel conditions in these trials are statistically as distant from the channel conditions of the first trial. Therefore, the extra degradation from the second to the third trial may be attributed to the time lapse aspects.

5. ADAPTATION TECHNIQUES

To see if the above theory is correct and in trying to alleviate the time lapse degradation, different adaptation techniques may be used. Adaptation techniques have been discussed in the literature and they mostly try to adapt a speaker's model to a Universal Background Model (**UBM**), see [3, 6, 17]. Here, we will further use adaptation to change the model for a candidate from the originally adapted model based on the first enrollment data to a new model which will be more resilient to changes in the channel and the time lapse effects. The first technique is data augmentation.

5.1. Data Augmentation

To modify the model for a speaker using data augmentation, the original enrollment data is retained for the candidate. At a point when a positive ID of the candidate is made, extra data is appended to the original enrollment data to provide a more universal enrollment model for the candidate matching different channel conditions and time lapse changes. Figures 3 and 4 show the identification and verification results respectively, using the new models enrolled by utilizing the augmented data. The results are in-tune with expectations.

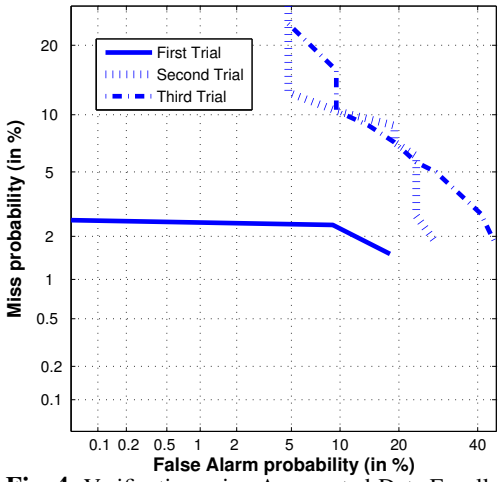

Fig. 4. Verification using Augmented-Data Enrollment

The identification performance degrades from a perfect performance to a 5% error rate. This is due to the contamination of the enrollment data which now contains channel information from trial 2 as well as trial 1. It also results, quite expectedly, in a large improvement for the second trial. This is the case, since now the model contains channel information from the second trial. However, there is also a very big improvement in the third trial which is partly attributed to the smoother information content about channels contained in the enrollment data, although no data from the third session was used in the enrollment. Since no specific information is contained in the this enrollment about the channel dynamics of the third trial, this smoothness is apparently attributing to better general time lapse performance.

Similar results are seen in figure 4 for verification of the second and third trials. In fact, comparing figures 2 and 4 shows that there is no degradation in the verification of the first trial with the EER still being around 2.5%. However, the EER of the second and the third trial have been reduced to only 10%,

though the overall performance in the second trial is still better than the third trial as expected.

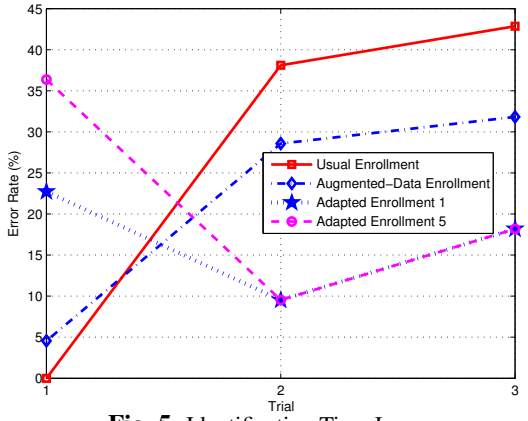

Fig. 5. Identification Time Lapse

5.2. Maximum A Posteriori Adaptation

One of the problems with the augmented-data approach of the previous section is that the original audio data has to be maintained to be able to do a re-enrollment by adapting from the speaker-independent model (the speaker model representing the general population) to the speaker model for the augmented data. Also, conceptually, the same weight is given to the old data as is given to the new data. One remedy is to use the adaptation techniques which were used to adapt from the speaker-independent model to the speaker model, to adapt from the speaker model to a new speaker model considering the new data at hand. The adaptation technique which was used here is the Maximum A-Posteriori adaptation method. Other techniques such as Maximum Likelihood Linear Regression (**MLLR**) may have very well been used for this purpose. [3, 17]

In doing the MAP adaptation, the number of iterations dictate the forgetting factor of the technique. The higher the number of iterations, the more the new data is considered in contrast to the old data. Normally, about 5 iterations are used to go from the speaker-independent model to the speaker model. Initially, this number was used to further transform the prototypes from the old model to the new model for the speaker.

Figure 5 shows the results for identification using the new models compared to the usual enrollment and the augmented-data enrollment. The MAP adapted enrollment using 5 iterations shows much better overall performance than both of the usual enrollment and augmented-data enrollment models. However, because it over-trains on the data of the second trial, the results of the first trial are highly degraded. To remedy this problem, the number of iterations for this MAP adaptation was reduced to 1. The results are shown in the same figure (5). The results show that no degradation is reported for the second and third trials, however, the identification performance of the first trial is greatly improved.

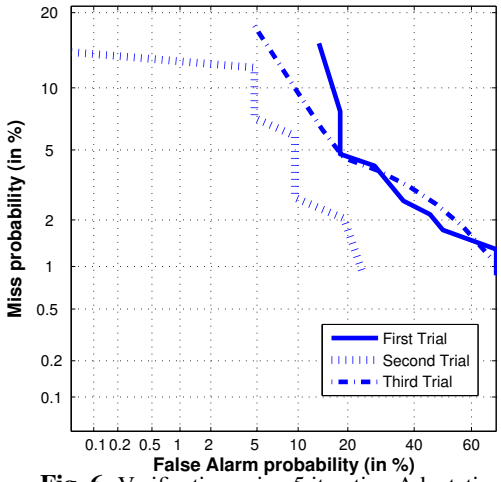

Fig. 6. Verification using 5 iteration Adaptation

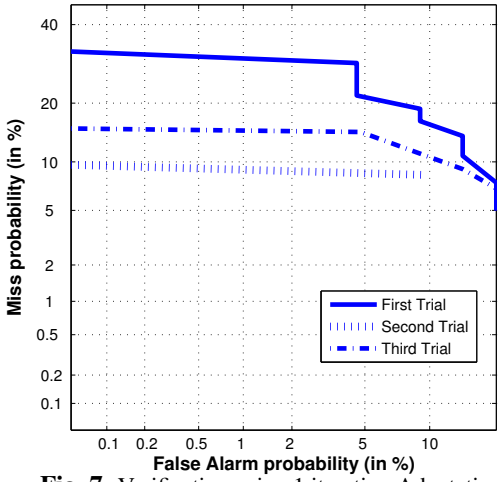

Fig. 7. Verification using 1 iteration Adaptation

Figures 6 and 7 show similar performance for the verification case. Again, using 5 iterations degrades the first trial by over-training. Figure 7 portrays much better performance across the different trials. In addition, the third trial has quite an acceptable performance although no channel information has been included in the speaker model from this trial. The EER for all three trials using a MAP adaptation with one iteration varies from about 10% for the best case which is trial two (the trial for which we have adapted) to a respectable maximum of about 14% for the worst case which is trial one from which we deviated. Trial three has an EER of about 12% which seems to coincide with the average EER across the different trials.

6. CONCLUSION

We have seen that there seem to be other effects in addition to channel-mismatch which further degrade the identification and verification performance of a statistical speaker recognition system. We have lumped all these effects into a category called time lapse effects. These effects along the side of channel-mismatch effects were somewhat suppressed using an augmented data approach where the enrollment audio data is always kept around and augmented with new data whenever a positive ID is made and this way the overall performance increases. Although this degrades the best case performance of the engine.

One of the problems with keeping the enrollment audio data is the memory-intensive nature of the solution. In addition security breaches may occur including legal and constitutional issues with keeping audio data around on a server. Some constitutions including that of the United States of America attach an ownership to the raw audio of a person. In addition, compromised access to the server holding the audio data will cause security breaches such as spoofing capabilities, etc. [3] In order to remedy these problems and the performance degradation issues, we used a MAP adaptation technique to adapt an existing speaker model to a new model using new enrollment data. It was shown that using non-aggressive adaptation works a lot better since over-training causes an overall degradation in the performance of both identification and verification engines. From the results we may further deduce that there is indeed a time-dependent degradation which may be remedied by using smoother models with more information across the time-line as well as different channels.

We have only scratched the surface of the time lapse issue and plan to do much further research in this area to do better speaker model smoothing using other compensation techniques such as MLLR [17] and Latent Factor Analysis (LFA) [5]. At the present, the study is being expanded to include over 100 speakers and to experiment with more re-takes to see how the time lapse effects and the adaptation results follow the trends seen here.

7. REFERENCES

[1] Francis Nolan, *The phonetic bases of speaker recognition*, Cambridge University Press, New York, 1983, ISBN: 0-521-24486-2.

[2] Naini A. S. and M. M. Homayounpour, "Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods," in *The 8th International Conference on Signal Processing*, 2006, vol. 1.

[3] Homayoon Beigi, *Fundamentals of Speaker Recognition*, Springer, New York, 2009, ISBN: 978-0-387-77591-3.

[4] Claude E. Shannon, "Communication in the presence of noise," *Proceedings of the Institute of Radio Engineers*, vol. 37, no. 1, pp. 10–21, Jan. 1949, Reprint available at: Proceedings of the IEEE, Vol. 86, No. 2, Feb. 1998.

[5] Robbie Vogt and Sridha Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech and Language*, vol. 22, no. 1, pp. 17–38, Jan. 2008.

[6] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[7] D.A. Reynolds, "Channel robust speaker verification via feature mapping," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, Apr 2003, vol. 2, pp. II–53–6.

[8] Teunen R., Shahshahani B., and Heck L., "A model-based transformational approach to robust speaker recognition," in *International Conference on Spoken Language Processing*, 2000, vol. 2, pp. 495–498.

[9] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 42–54, 2000.

[10] C. Barras and J.-L. and Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, Apr 2003, vol. 2, pp. II–49–52.

[11] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey - The Speaker Recognition Workshop*, Jun 2001, pp. 213–218.

[12] S. Gurbuz, J.N. Gowdy, and Z. and Tufekci, "Speech spectrogram based model adaptation for speaker identification," in *Southeastcon 2000. Proceedings of the IEEE*, Apr 2000, pp. 110–115.

[13] E. Endres, W. Bambach, and G. Flösser, "Voice spectrograms as a function of age, voice disguise and voice imitation," *Journal of the Acoustical Society of America (JASA)*, vol. 49, pp. 1842–1848, 1971.

[14] ITU-T, "G.711 Pulse Code Modulation (PCM) of Voice Frequencies," ITU-T Recommendation, Nov. 1988.

[15] Stefan Meltzer and Gerald Moser, "Mpeg-4 he-aac v2 – audio coding for today's digital media world," World Wide Web, 2005.

[16] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Eurospeech 1997*, 1997, pp. 1–8.

[17] Sungjoo Ahn, Sunmee Kang, and Hanseok Ko, "Effective speaker adaptations for speaker verification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2000)*, Jun 2000, vol. 2, pp. 1081–1084.