Addressing Sampling-Frequency Mismatch between Speech Data Sets in a Forensic Voice Comparison

Hanie Mehdinezhad[1], *Bernard J. Guillemin*[1], *Balamurali B T*[2]

[1]The University of Auckland, New Zealand
[2]Singapore University of Technology & Design, Singapore
E-mail: *h.mehdinezhad@auckland.ac.nz*
E-mail: *b.guillemin@auckland.ac.nz*
E-mail: *balamurali_bt@sutd.edu.sg*

Abstract

Sampling-frequency (f_s) mismatch between Suspect, Offender, and Background speech data sets in a Forensic Voice Comparison (FVC) are discussed and approaches to correct for this are presented. The Bayesian Likelihood-Ratio (LR) framework is used to express the results of a FVC and Gaussian Mixture Model-Universal Background Model (GMM-UBM) is used to calculate LR values. As appropriate, experiments have been conducted on both tokenized and stream data using Mel-Frequency Cepstral Coefficients (MFCCs) as the speech features. The results show that the best approach to correct for f_s-mismatch between speech data sets is down-sampling of the speech data set/sets at higher f_s to match the speech data set/sets at lower f_s.

Index Terms: Sampling frequency mismatch, Forensic Voice Comparison, Bayesian Likelihood Ratio, Gaussian Mixture Model-Universal Background Model.

1. Introduction

In a typical Forensic Voice Comparison (FVC) scenario, the similarity of an Offender voice sample to a Suspect voice sample and the typicality of an Offender voice sample in reference to a Background population, which is case-specific, are assessed to provide the strength of speech evidence. The results of a FVC are then used in order to assist a court of law to make their final decision. In a real forensic scenario, the Offender speech data could be obtained in a number of ways where there is often little/no control over it. For instance, Offender speech data could be obtained from the recordings of a mobile phone or landline phone conversation. On the other hand, there is normally more control over the speech for the Suspect (often recorded during a police interview) and the Background population (usually recorded under highly controlled recording conditions).

In a real forensic investigation, it could be possible that the speech samples associated with Suspect, Offender, and Background data are not at the same sampling-frequency (f_s). For example, the Offender data may be at a lower f_s compared to that of the Suspect and Background data. But when undertaking automatic-based FVC procedures using, for example, MFCCs, all speech data sets must be at the same f_s. Therefore, it is necessary to determine what strategy should be employed in cases where, for example, the Offender speech data is sampled at a different f_s compared to Suspect and Background speech data. Hence, the focus of this paper is on answering the following research question: If the voice samples associated with the Suspect, Offender, and Background data sets are not at the same f_s (this is a form of mismatch that can be called f_s-mismatch between speech data sets), what strategy should be employed to make them the same? We consider the specific case of Offender data being at a lower f_s compared to that of the Suspect and Background data. Should it be up-sampled to match them, or conversely should the Suspect and Background data be down-sampled?

The Bayesian Likelihood Ratio (LR) framework is used to express the outcome of a FVC and a Gaussian Mixture Model-Universal Background Model (GMM-UBM) is used as the statistical procedure to calculate LRs. The speech database used in this study is the XM2VTS (Extended Multi Modal Verification for Teleservices and Security) that is a multi-modal database containing speech recordings sampled at 32 kHz. Thus, 32 kHz is the highest f_s that could be investigated in this research. Mel-Frequency Cepstral Coefficients (MFCCs) are the speech features used as they are commonly used in the FVC arena.

The remainder of this paper is organized as follows. Section 2 provides an overview of the LR framework, followed by a brief discussion on GMM-UBM. Section 3 describes our experimental procedures for addressing f_s-mismatch between speech data sets. The results of these experiments are presented in Section 4, followed by our conclusions in Section 5.

2. Background Information

2.1. *Likelihood Ratio Framework*

The LR, as the name indicates, is the ratio of two likelihoods that mathematically is calculated as:

$$LR = \frac{P(E|H_p)}{P(E|H_d)} \qquad (1),$$

where $P(E|H_p)$ is the conditional probability of E (the evidence) given H_p (the prosecution hypothesis) and this assesses the similarity between the Suspect and Offender speech samples. $P(E|H_d)$ is the conditional probability of E given H_d (the defense hypothesis) and measures the typicality of the Offender speech samples to a relevant Background population. LR values significantly greater than one support the prosecution hypothesis, LR values significantly less than one support the defense hypothesis, and LR values close to one provide little support either way. The Log-Likelihood-Ratio (LLR) is often computed from the LR, where $LLR = log_{10}(LR)$. The sign of the LLR indicates whether it supports the prosecution (positive) or defense (negative) hypothesis and its magnitude indicates the strength of that support.

2.2. *Overview of GMM-UBM*

The GMM-UBM approach [1] is a common technique used in both Automatic Speech Recognition (ASR) and FVC [2-4]. Normally it requires a large amount of data to build a single Background model, namely a Universal Background Model (UBM) [2]. GMM-UBM was originally designed for data-stream-based analysis, but it has also been used by some researchers in token-based analysis [4]. In order to achieve good FVC performance, the UBM is trained on all Background data pooled across speakers. The probability density function of the UBM is estimated using Gaussian Mixture Models (GMMs), with the Expectation Maximization (EM) algorithm [5, 6] being used to train it. The Suspect model is then built by adapting the UBM towards a better fit of the Suspect speech data using the Maximum a Posteriori (MAP) procedure [2]. A score is then calculated as the ratio of the Suspect and Background probability density function values determined at the Offender data points. The scores are then calibrated and fused to get the LR.

2.3. *Measuring FVC performance / Presenting results*

The performance of a FVC is measured by evaluating its accuracy (i.e., validity) and reliability (i.e., precision) [7, 8]. Accuracy indicates the closeness of the obtained result to the true value of the output. The Log-Likelihood Ratio Cost (C_{llr}) [8, 9] is one of the recommended metrics for assessing this, the lower its value, the better the accuracy. Reliability measures the amount of variation that could be expected in LR values, this arising, for example, from improper modelling of the background statistics due to limited data being available of the specified Background population. The Credible Interval (CI) [8, 9] is a popular metric for evaluating this, and again, the lower its value, the better the reliability.

The results of a FVC experiment are often presented using Tippett plots [10] which represent the cumulative proportion of LLR values for both same-speaker and different-speaker comparisons. In these plots (see Figs. 2 & 3) the solid blue and solid red curves are the same-speaker and different-speaker comparison results, respectively. Since positive LLR values support the prosecution hypothesis and negative values support the defense hypothesis, the further apart the curves (i.e., the blue curve towards the right and the red curve to the left), the better would be the performance of the FVC system and therefore generally the lower the C_{llr}. The dashed lines on either side of these solid curves represent the variation in a particular LLR comparison result (i.e., $LLR \pm CI$). The lower the CI value, the higher the reliability of the FVC system.

3. Experimental Procedure

3.1. *Speech Data Set*

The XM2VTS database includes video sequences, face images, and speech recordings [11]. It contains read speech digitized at 16 bits, sampled at 32 kHz, with low Background Noise (BN) level. The language in the database is English with predominantly a Southern British accent. It contains four recording sessions of 295 subjects (156 males, 139 females) collected over a period of 4 months. Sessions were recorded at one-month intervals and during each session each speaker repeated three sequences of words twice. The first two were random sequences of digits from zero to nine: “zero one two three four five six seven eight nine” and “five zero six nine two eight one three seven four”. The last sequence was a sentence: “Joe took father’s green shoe bench out”. In this research only the recordings of the first two sentences (i.e., random sequences of digits) have been used.

Given that the XM2VTS database contains recordings of read speech and the BN level is low, it is acknowledged that it is not very forensically realistic. However, in support of its use in this study, it does include a large number of speakers with similar accents as well as multiple non-contemporaneous recordings, both aspects being highly important in the FVC arena. Only male speaker recordings have been used in this study. Of the 156 male speakers, only 130 were used. The other 26 speakers were discarded because their recordings were either less audible, or they were judged to have different accents to the rest of the speakers (see [12] for a rationale behind discarding recordings on the basis of dissimilar accent).

For our experiments we then down-sampled the recordings to 8 kHz and then up-sampled again depending upon the scenario we wished to investigate. For data-stream-based experiments, the whole utterance (i.e., the two random sequences of digits) after removing all silence segments was used. For the token-based experiments, two diphthongs /aɪ/ and /eɪ/ and one monophthong /iː/ extracted from the words “nine”, “eight” and “three”, respectively, were used. Audio editing programs Wavesurfer [13] and Goldwave [14] were used to assist the extraction process, as well as the processes of down-sampling and up-sampling.

MFCCs have been shown to provide good comparison performance in the FVC arena [4, 15]. To calculate MFCCs, first the speech signal is converted into the frequency domain using the Discrete Fourier Transform (DFT). Then the energies in various regions of the frequency domain are estimated over a set of overlapped Mel-filter banks. MFCCs are then calculated by taking a Discrete Cosine Transform (DCT) of the logarithm of energies. Different numbers of MFCCs (typically, between 12 to 16) have been used by researchers [16, 17]. At a very early stage of this study, some FVC experiments were undertaken using different numbers of MFCCs. Based on the results of those experiments, it was decided to use 14 MFCCs in our token-based experiments. To take advantage of the delta-Cepstral feature, 14 deltas and 14 delta-deltas were added to the 14 MFCCs in the data-stream-based analysis.

3.2. *Comparison Process*

The 130 male speakers were divided into three mutually exclusive sets: 44 speakers for the Background set and 43 speakers each for the Development and Testing sets. (Note: the FVC results from the Development set are used to calibrate and fuse the results from the Testing set [9]). Data from three of the four recording sessions were used for the speakers in the Background set, while all four recording sessions were used for each of the speakers in the Development and Testing sets. The Suspect model for each comparison was formed using data from recording Sessions 1 and 2. This gives eight tokens per vowel for token-based analysis and eight utterance-segments per speaker for data-stream-based analysis. Sessions 2, 3 and 4 were used in turn for the Offender data. For the same-speaker comparisons, Sessions 3 and 4 for each speaker (i.e., Offender data) were compared with the Suspect model of the same speaker. For the different-speaker comparisons, Sessions 2, 3 and 4 for each speaker (i.e., Offender data) were compared with the Suspect models for the other speakers. More details of these comparisons can be found in [18].

With 43 speakers in each of the Testing and Development sets, 43 same-speaker comparisons and 903 different-speaker comparisons are possible (ignoring multiple comparisons required in order to compute the CI). The results for individual vowels for the token-based analyses were then calibrated and fused, but for data-stream analyses, the results were just calibrated. Calibration and fusion were achieved using logistic regression [9]. The C_{llr} was calculated from the average of LRs for the two same-speaker comparisons and the average of LRs for the three different-speaker comparisons. The CI for both same-speaker and different-speaker comparison results were computed using the procedure outlined in [8].

3.3. *Experimental Set-up for Addressing f_s-Mismatch between Suspect, Offender, and Background Data*

To address f_s-mismatch between speech data sets in a FVC experiment, three scenarios have been investigated: (a) the Offender data sampled at 8 kHz, with the Suspect and Background data sampled at 32 kHz. The Suspect and Background data were then down-sampled to 8 kHz to match the Offender data, (b) the Offender data originally at 8 kHz and then up-sampled to match the Suspect and Background data at 32 kHz, and (c) the Offender, Suspect, and Background data originally at 8 kHz (having previously been down-sampled from 32 kHz) and then up-sampled back 32 kHz. It is acknowledged that this last scenario is not very realistic in the sense of why would one want to do that in a real forensic case, but this experiment was carried out for the sake of completeness to see whether it produced any unexpected results. As an example, the block diagram for the set-up of Scenario (b) is shown in Fig. 1.

4. Results

Table 1 shows the mean C_{llr} and 95% CI for Scenarios (a) – (c). It is clear that Scenario (b) results in very poor FVC performance, as can be seen from mean C_{llr} values being close to 1 for token-based analysis and relatively high for data-stream-based analysis.

Table 1: *Mean C_{llr} and CI for investigations that addressed f_s-mismatch between speech data sets in a FVC*

Sampling-Frequencies		Token-based	Data-stream-based
Scenario (a): Offender Speech: 8 kHz Suspect & Background Speech: 32 kHz to 8 kHz	Mean C_{llr}	**0.214**	**0.008**
	95% CI	2.245	1.069
Scenario (b): Offender Speech: 8 kHz to 32 kHz Suspect & Background Speech: 32 kHz	Mean C_{llr}	**0.903**	0.796
	95% CI	0.561	0.270
Scenario (c): Offender Speech: 8 kHz to 32 kHz Suspect & Background Speech: 8 kHz to 32 kHz	Mean C_{llr}	**0.282**	**0.021**
	95% CI	3.485	2.257

Comparing Scenarios (a) and (c), Scenario (c) has clearly resulted in a worse performance than Scenario (a) for both token-based and data-stream-based analyses. But the degradation in performance is certainly not as severe as that for Scenario (b).

So, the obvious question is: What is it about Scenario (b) that has resulted in such poor FVC performance? In Scenario (a), the Offender, Suspect, and Background data sets have all been down-sampled to 8 kHz. The resulting loss of spectral information in the frequency band 4-16 kHz for all three data sets would thus be the same. (Note: with sampled signals it is only possible to retain spectral information up to half the f_s.) But with Scenario (b), the Offender data was first down-sampled to 8 kHz, thereby losing spectral information in the band 4-16 kHz, then up-sampled back to 32 kHz. But even though it has been up-sampled back to 32 kHz, the spectral information it lost in the 4-16 kHz band as a result of down-sampling remains lost. However, the Suspect and Background data remained at 32 kHz and thus suffered no loss of spectral information in the 4-16 kHz band. So, one is no longer comparing 'like with like' with Scenario (b). Specifically, the up-sampled Offender data contained speech spectral energy only in the band 0-4 kHz, whereas the Suspect and Background data contained speech spectral energy in the band 0-16 kHz. We hypothesize that this is the reason for the very poor FVC performance of Scenario (b).

Returning again to a comparison of Scenarios (a) and (c), it is surprising that the FVC performance for Scenario (c) is somewhat worse than for Scenario (a). For Scenario (a), all three data sets were down-sampled to 8 kHz, so they then contained speech spectral energy in the band 0-4 kHz. Though in Scenario (c) all three data sets were first down sampled to 8 kHz and then up-sampled back to 32 kHz, they still only contained speech spectral energy in the band 0-4 kHz. So, in both scenarios one should be comparing 'like-with-like'. One expects this up-sampling process to be reasonably transparent in terms of not losing or degrading information. But the results here suggest that in the FVC arena this is clearly not the case.

A final observation can be gleaned from Table 1, namely that data-stream-based analysis outperforms token-based analysis in all scenarios. Given that the former utilises much more information, this is to be expected.

Figs. 2 and 3 show Tippett plots for Scenarios (a) to (c) for token-based and data-stream-based analysis, respectively. Considering first the plots of Figs. 2 and 3 for Scenario (b), it is clear that both same- and different-speaker comparisons have been similarly and adversely affected, and there is a large number of misclassifications for both. In addition, both the red line (representing different-speaker comparisons) and the blue line (representing same-speaker comparisons) are very close to the green line (i.e., LLR=0). This means that the LLR (or LR) is not providing much useful speaker-specific information, and this is true for token- and stream-based analysis.

Comparing now Scenarios (a) and (c) in Figs 2 and 3, it can be seen that for token-based analysis there is an increase in the number of same-speaker misclassifications for Scenario (c) compared to Scenario (a). For data-stream-based analysis, there are no same-speaker misclassifications for either scenario, but the performance of same-speaker comparisons for Scenario (a) is better than for Scenario (c). In terms of reliability, it can be seen that the 95% CI in Scenario (c) for both same- and different-speaker comparisons is larger than it is for Scenario (a).

5. Conclusions

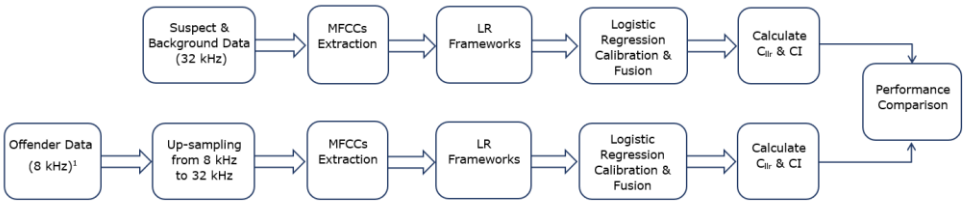
This paper has addressed f_s-mismatch between Suspect, Offender, and Background data sets in a FVC and what strategy should be employed to correct for this. The LR framework has been used for the experiments, and GMM-UBM has been used to compute LR values both for token-based and data-stream-base analyses.

As the XM2VTS, the speech database used in our study, contains speech data sampled at 32 kHz, this is the highest f_s used in our experiments. Speech data was then down-sampled to 8 kHz and then up-sampled again depending upon the scenario to be investigated. MFCCs were the speech features used in our investigations.

Three scenarios have been investigated: (a) the Offender data at 8 kHz and the Suspect and Background data down-sampled to 8 kHz to match, (b) the Offender data up-sampled from 8 kHz to 32 kHz to match the Suspect and Background data at their original 32 kHz, and (c) the Suspect, Offender, and Background data all up-sampled from 8 kHz to 32 kHz, having previously been down-sampled from 32 kHz to 8 kHz.

The results of our experiments conclusively show that Scenario (a) is the strategy to be adopted. The converse, i.e., Scenario (b), results in exceedingly poor FVC performance because, we hypothesise, one is then not comparing 'like-with-like'. Our results have also shown that up-sampling of all data sets from 8 kHz to 32 kHz, i.e., Scenario (c), results in poorer FVC performance, though not significantly, than leaving them all at 8 kHz. So, we conclude that there is no advantage in doing this.

The speech database used in this research is not very forensically realistic as it contains good quality speech recordings. When using more forensically-realistic data, we expect to get somewhat poorer results, but nonetheless, we do expect to observe the same trend. The reason is that our observed trend is related to a fundamental issue (i.e., spectral-data mismatch) and not the quality of the speech data.



[1] Offender data obtained from MX2VTS speech database which is originally sampled at 32 kHz. For this experiment, it was first down-sampled to 8 kHz.

Figure 1: *Experimental set-up for addressing f_s-mismatch between speech data sets in a FVC (Scenario (b))*

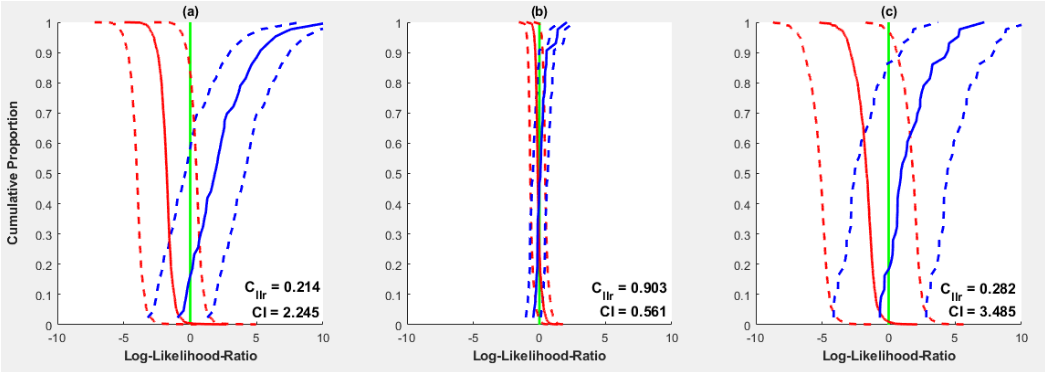


Figure 2: *Tippett plots showing the performance of GMM-UBM when applied to token-data: (a) Offender speech data at 8 kHz, Suspect and Background speech data are down-sampled from 32 kHz to 8 kHz; (b) Background and Suspect speech data are originally at 32 kHz, Offender speech data are up-sampled from 8 kHz to 32 kHz; (c) all three data sets are up-sampled from 8 kHz to 32 kHz*

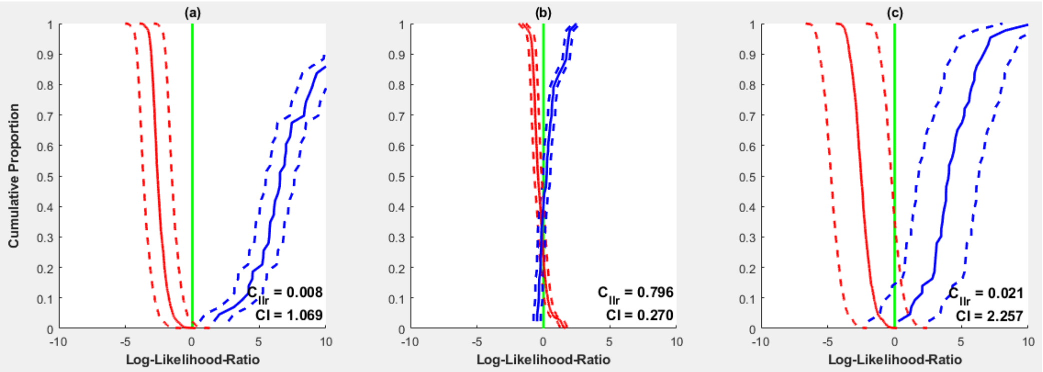


Figure 3: *Tippett plots showing the performance of GMM-UBM when applied to stream-data: (a) Offender speech data at 8 kHz, Suspect and Background speech data are down-sampled from 32 kHz to 8 kHz; (b) Background and Suspect speech data are originally at 32 kHz, Offender speech data are up-sampled from 8 kHz to 32 kHz; (c) all three data sets are up-sampled from 8 kHz to 32 kHz*

6. References

1. Reynolds, D., *Gaussian Mixture Models,* . Encyclopedia of Biometric Recognition, Springer, 2008.
2. Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker verification using adapted Gaussian mixture models.* Digital signal processing, 2000. **10**(1): p. 19-41.
3. Reynolds, D.A. *Automatic speaker recognition using Gaussian mixture speaker models*. in *The Lincoln Laboratory Journal*. 1995. Citeseer.
4. Morrison, G.S., *A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM).* Speech Communication, 2011. **53**(2): p. 242-256.
5. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 1977. **39**(1): p. 1-22.
6. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. 2009: Springer Science & Business Media.
7. Morrison, G.S., *Forensic voice comparison and the paradigm shift.* Science & Justice, 2009. **49**(4): p. 298-308.
8. Morrison, G.S., *Measuring the validity and reliability of forensic likelihood-ratio systems.* Science & Justice, 2011. **51**(3): p. 91-98.
9. Morrison, G.S., *Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio.* Australian Journal of Forensic Sciences, 2013. **45**(2): p. 173-197.
10. Meuwly, D. and A. Drygajlo. *Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)*. in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*. 2001.
11. Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999. Citeseer.
12. Morrison, G.S., P. Rose, and C. Zhang, *Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice.* Australian Journal of Forensic Sciences, 2012. **44**(2): p. 155-167.
13. Wavesurfer, *Wavesurfer. (2012, December 5 – last update). Retrieved on 12 January 2013, last retrieved from http://www.speech.kth.se/wavesurfer/index2.html*. 2011.
14. GoldWave, *GoldWave. (2013, January 24 – last update). Retrieved on 5 March 2013, last retrieved from http://www.goldwave.com/*. 2012.
15. Nair, B.B., E.A. Alzqhoul, and B.J. Guillemin. *Comparison between Mel-frequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework*. in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*. 2014.
16. Rabiner, L.R. and R.W. Schafer, *Introduction to digital speech processing*. 2007: Now Publishers Inc.
17. Vandyke, D.J., *Glottal Waveforms for Speaker Inference & A Regression Score Post-Processing Method Applicable to General Classification Problems*. 2014, Citeseer.
18. Mehdinezhad, H. and B.J. Guillemin, *Preliminary performance comparison between PCAKLR and GMM-UBM for computing the strength of speech evidence in forensic voice comparison.* SST2016, 16th Speech and Science Technology, 2016.