Comparison between Speech Parameters for Forensic Voice Comparison Using Mobile Phone Speech

Esam A. S. Alzqhoul [1, 2], *Balamurali B. T. Nair* [1, 2], *Bernard J. Guillemin* [1, 2]

[1] Forensic and Biometrics Research Group (FaB), The University of Auckland, New Zealand
[2] Department of Electrical and Computer Engineering, The University of Auckland, New Zealand
ealz002@aucklanduni.ac.nz, bbah005@aucklanduni.ac.nz, bj.guillemin@auckland.ac.nz

Abstract

Amongst the various speech parameters available for forensic voice comparison (FVC), Mel-frequency cepstral coefficients (MFCCs) have been found to give good performance and are widely used. The aspect we address in this paper is whether they are also the optimum choice for the specific case of mobile phone speech recordings. We have included a number of cepstral coefficient parameters sets in our investigation, including MFCCs, as well as formant trajectories, these being a more traditional parameter set used by forensic speech scientists. Our results suggest that MFCCs are also the optimum choice for mobile phone speech.

Index Terms: forensic voice comparison, mobile phone speech, optimum parameter set, cepstral coefficients

1. Introduction

Analysis of speech recordings can play an important role in assisting a court in determining the identity of an offender. These recordings vary in their quality depending on their origin, which can be a police interview room, landline or mobile phone network, etc. Various parameters can be extracted from the speech signal to calculate the strength of evidence. Some of the typically used parameters are cepstral coefficients, specifically Mel-frequency cepstral coefficients (MFCCs), and formants. This paper investigates the Forensic Voice Comparison (FVC) performance of various types of cepstral coefficients, as well as speech formants, using mobile phone speech recordings.

There are a number of technologies in the mobile phone arena such as Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA) and these are very different in their ways of handling, processing and transmitting the speech signal [1]. The mobile speech recordings used in this investigation have not been transmitted through an actual mobile phone network. One could do so, but this will reveal only a small subset of the actual transmission scenarios in the network. In our view a better strategy is to use the respective speech codec of these networks to code the speech samples under various modes of operation. This is because the speech codec in these networks is the only component that is directly responsible for the quality of the resulting transmitted speech signal [2]. Other factors in the network such as poor channel conditions and channel noise do not impact directly on the speech signal, but rather indirectly by way of instructions sent to the codec from upper levels of the network. These instructions change the codec modes of operation to minimize the impact of these external factors. The most widely used speech codecs in the GSM and CDMA mobile networks are the Adaptive Multi Rate Codec (AMR) [3] and Enhanced Variable Rate Codec (EVRC) [4], respectively.

The Bayesian likelihood-ratio (LR) framework has been used in our experiments to quantify the strength of speech evidence and Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) [5] has been chosen for computing LRs. This choice is motivated by the fact that PCAKLR can handle a large number of speech parameters with no restriction on the amount of data required. It has also been found to provide comparable results to the Multivariate Kernel Density (MVKD) analysis when used with a small number of parameters [5]. The performance of FVC analysis has been estimated using the following tools: log-likelihood-ratio cost (C_{llr}), and Credible Interval (CI) and results are shown using Tippett plots [6-10].

The remainder of this paper is structured as follows. The key differences between the GSM and CDMA networks, together with their associated codecs, are discussed in Section 2. Section 3 overviews the various speech parameters used in our experiments. Section 4 discusses the experimental methodology chosen for this investigation. Results and findings are presented in Section 5 and conclusions and comments are presented in Section 6.

2. Speech codecs in the GSM and CDMA networks

The key difference between the GSM and CDMA networks is in respect to the mechanisms determining the quality of the transmitted speech signal. With the GSM network it is the changes in channel quality which largely control this, whereas in the CDMA network it is the changes in channel capacity. The codecs in these networks are instructed by the network to account for these changing factors.

The AMR codec employed in the GSM network uses a speech coding technique called Algebraic Code-excited Linear Prediction (ACELP). It can operate at one of eight source coding bit rates: 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20 and 12.20 kbps [3]. The codec switches between these bit rates as often as every 40ms upon receiving instruction from the network [11]. The advantage of the AMR codec is its ability to create a balance between the speech coding bits and error protection bits transmitted. This balance changes dynamically in response to changing channel quality [2].

The EVRC used in the CDMA network incorporates a number of speech coding techniques such as Code-excited Linear Prediction (CELP), Noise Excited Linear Prediction (NELP) and Pitch Period Prototype (PPP). The speech signal can be coded at one of four source coding bit rates: 0.8, 2, 4 and 8.55 kbps. Unlike the AMR codec, the EVRC can change the bit rate as often as every 20ms. The EVRC is instructed by the network to change its mode of operation when the channel interference levels increase as a result of increasing number of users accessing the system simultaneously [1, 4].

3. Speech Parameters

3.1. Cepstral Analysis

Cepstral analysis is one of the processes used to separate aspects of the source and filter components of speech [12]. This type of analysis, often referred to as homomorphic filtering, is not new and it has been used in other fields [13]. Cepstral analysis is generally defined as the inverse Fourier transform of the logarithm of the Fourier transform of a signal. When applied to speech, the Fourier transform is used to convert the convolution in the time domain between the major parts of the source-filter model (i.e., excitation, glottal shaping, vocal tract and lip radiation filters) into a product of their corresponding representations in the frequency domain. The logarithm operator is then used to transform the product into an addition of components. The inverse Fourier transform is then applied to bring the separated components back into the time domain, referred to as the quefrency domain [12].

3.1.1. Complex Cepstral Coefficients (CCCs)

The complex cepstral coefficients (CCCs) can be extracted using the standard cepstral analysis procedure as mentioned above. They carry information about the magnitude and phase of all components of the source-filter model. The CCCs are always real because the poles and zeroes of the speech production model are either real or occur in complex conjugate pairs. The CCCs contain both causal and anti-causal components, the former arising from poles and zeroes inside the unit circle in the z-plane that belong to the vocal tract and lip radiation filters. The anti-causal part originates from the glottal shaping filter as a result of zeroes outside the unit circle [12].

3.1.2. Real Cepstral Coefficients (RCCs)

The process of extracting real cepstral coefficients (RCCs) differs slightly from the one used for CCCs. Specifically, the logarithm is applied to the magnitude of the Fourier transform and phase information is ignored. Thus the RCCs are the even component of the CCCs and therefore inherently carry less information.

3.1.3. Linear Prediction Cepstral Coefficients (LPCCs)

The linear prediction cepstral coefficients (LPCCs) are derived directly from the LPC coefficients. LPC analysis assumes a simplified source-filter model which lumps together vocal tract, lip radiation and glottal shaping filters and assumes that the resultant filter is all-pole with poles inside the unit circle. Consequently the LPCCs contain causal components only [14]. The lumped nature of the LPC model, together with the fact that zeroes in the speech signal are ignored, suggest that LPCCs do not carry as much information as either the CCCs or RCCs and could be expected, therefore, to perform somewhat worse in FVC.

3.1.4. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are a perceptual-based parameter. The speech signal is first converted to the frequency domain using the Discrete Fourier Transform (DFT). The next step is estimating how much energy exists in various regions of the frequency domain. This is estimated over a set of overlapped Mel-filter banks by computing the power spectrum of the speech signal and then summing up the energies in each filter bank region. Once the filter bank energies are computed, the logarithm operator is applied. Finally, a Discrete Cosine Transform (DCT) on the logarithm of the energies is performed resulting in a set of MFCC coefficients [12]. Though, the general process of MFCC extraction aligns with the procedure of cepstral analysis, it is still arguable whether MFCCs can be classified as a type of cepstral analysis or not. This is because MFCCs do not involve a separation of the speech components, as is the case for other types of cepstral coefficients.

3.2. Formant Analysis

Speech formants are identified by locating the dominant peaks of the speech spectral envelope [15]. Changes in formant values from one speech frame to another constitute the formant trajectory. Formant trajectories capture the dynamic aspects of speech and are widely used in speech applications such as speech and speaker recognition and vowel separation. In the following experiments formants have been extracted by the process mentioned in [16]. Only the first three formants were considered in our experiments. Discrete cosine transforms were fitted to the formant trajectories and only the first four DCT coefficients were used for each of the formant trajectories.

4. Experimental Setup

4.1. The speech database

We have used the XM2VTS database in our experiments, which contains four non-contemporaneous recording sessions for each speaker separated by one month intervals. The speakers were asked to read a sequence of digits in a randomized manner. 130 male speakers in this database were considered here [17]. For each speaker three vowel segments, namely /aI/, /eI/ and /i/, were extracted from the words “nine”, “eight” and “three”, respectively. Out of the four sessions available, three of them have been used in the following experiments. Speech samples were down-sampled to 8 kHz and stored as 16 bit PCM wave files to align with the input speech requirements of both mobile codecs. The down-sampled speech was then coded under different modes of operation for both the GSM and CDMA networks (see Section 4.2). The 130 speakers were divided into three groups: 44 speakers for the Background set, 43 speakers for the Development set and 43 speakers for the Testing set. With 43 speakers in the Testing set, 43 same-speaker comparisons and 903 different-speaker comparisons are possible for one session.

4.2. Experimental methodology

Figure 1 shows a block diagram of our experimental procedure. This involves a performance comparison using C_{llr} and CI for various FVC analyses using different speech parameters. Each of these parameters has been extracted from the whole vowel segment, except for formant trajectories. This is because in our experience cepstral coefficients computed on a frame-by-frame basis exhibit poorer FVC performance in terms of precision than those extracted from an entire speech segment, even if the segment is non-stationary. The number of cepstral coefficients used in our experiments was 100 (i.e., 50 Causal and 50 anti-causal components), 50, 23 and 16 for CCCs, RCCs, MFCCs and LPCCs, respectively. This choice of 100 CCCs and 50 RCCs has been motivated by the fact that

the vocal tract components are dominant in this region of the cepstral domain [12]. The choice of 23 MFCCs and 16 LPCCs was motivated by our own experiments with mobile phone speech which have shown that these numbers of coefficients generally give optimum FVC performance. Finally, with 3 formant trajectories and 4 DCTs fitted for each of them, a total of 12 DCT coefficients were used for our formant trajectory experiments.

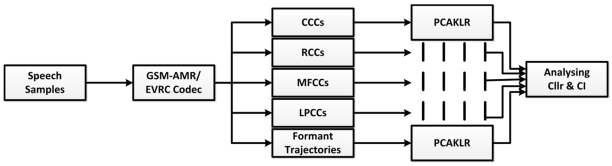

Figure 1: Block diagram of experimental procedure

The experiments have been repeated for two coding modes of each codec, these being chosen to correspond to high and low speech coding quality. The reason for this is that we are interested to know how much variation in the performance of an FVC analysis one could expect for each set of speech parameters as a function of speech coding quality. The coding bit rates chosen for the GSM codec experiments were 4.75 and 12.2 kbps, whereas anchor operating points OP2 (low quality) and OP0 (high quality) were used for the CDMA experiments. The choice of different modes in the EVRC was also motivated by the fact that each anchor operating point incorporates a different mix of the CELP, NELP and PPP coding techniques, and these may well have different impacts on the speech parameters of interest. It should be noted here that the Background set for both mobile phone network experiments were coded using their respective speech codecs.

5. Results

Two same-speaker comparison results were obtained for each speaker in the Testing set by comparing their Session 1 recording with their own recordings from Sessions 2 and 3. In respect to the different-speaker comparisons, three LRs were produced for each speaker by comparing their Session 1 recording with other speakers' recordings from Sessions 1, 2 and 3. For each comparison, the resulting scores from three vowel segments were combined using the standard logistic-regression fusion [18]. A mean LR value was then calculated using two LR estimates for each same-speaker comparison and three LRs for each different-speaker comparison. The accuracy of the resulting mean LRs has been computed using C_{llr}. The CI was estimated for each comparison and the reliability has then been expressed in terms of the average of these CI values. The final results are plotted using Tippett plots.

The performance of FVC analyses estimated in respect to the various speech parameters using low and high quality GSM and CDMA coded speech are shown in Tables 1 and 2, respectively. Generally speaking, for both networks, the accuracy of the FVC is better for MFCCs compared to other types of parameters. CCCs come second in the list followed by RCCs, LPCCs and finally formant trajectories. The accuracy of a FVC analysis using AMR coded speech improves with the bit rate. This is true for all cepstral coefficients except for LPCCs. For the CDMA experiments, the accuracy improves marginally with higher speech coding quality in the case of MFCCs and LPCCs. In respect to the formant trajectory experiments, accuracy surprisingly gets worse with higher quality of speech coding. The reasons for this are not clear, though the trend is consistent for both speech codecs.

Table 1: *Performance of speech parameters using GSM coded speech*

Parameters	4.75 kbps		12.2 kbps	
	C_{llr}	CI	C_{llr}	CI
CCCs	0.156	1.705	0.125	2.509
RCCs	0.199	1.633	0.132	2.258
MFCCs	0.141	1.568	0.108	1.603
LPCCs	0.215	1.554	0.231	2.186
Formant trajectories	0.341	1.423	0.442	1.479

Table 2: *Performance of speech parameters using CDMA coded speech*

Parameters	OP2		OP0	
	C_{llr}	CI	C_{llr}	CI
CCCs	0.106	2.048	0.135	2.265
RCCs	0.135	1.976	0.151	2.104
MFCCs	0.127	2.036	0.122	1.861
LPCCs	0.285	1.674	0.249	1.644
Formant trajectories	0.373	1.919	0.401	1.304

In terms of the reliability of LR results, the CI is much worse for most speech parameters (MFCCs and formant trajectories being the exception) when high quality GSM speech is used. In contrast, the CI improves for most of the speech parameters when higher speech coding quality is used in the CDMA network. Although, CCCs and RCCs carry more information about the glottal shaping and lip radiation filters, they surprisingly have not performed as well as MFCCs. The degradation in their accuracy is likely to be linked to the coding processes used in the AMR and EVRC codecs. These codecs have been designed to achieve the best speech quality at reasonably low bit rates. From the perspective of the codec designers, the glottal and lip radiation filters add only a little improvement to speech quality and thus these have not been considered in the codec design. As a result, any speech parameter related to these components would lose its discriminative power when coded.

The accuracy of FVC results was further investigated using Tippett plots. Figures 2 and 3 show the respective performance of CCCs and MFCCs when using high quality AMR speech coded at 12.2 kbps. The EVRC coded speech results coded at OP0 are shown in Figures 4 and 5. Figure 2 (CCCs) and Figure 3 (MFCCs) show similar performance in terms of different-speaker comparisons. However, the same-speaker misclassifications are marginally lower for MFCCs compared to CCCs. This explains the marginal improvement of accuracy when using the former. The Tippett plots of CCCs and MFCCs for the EVRC experiments (Figures 4 and 5) show comparable performance in terms of both the same- and different-speaker results.

6. Conclusions

MFCCs have been found to be the best performing parameter with mobile phone coded speech irrespective of the network being used. CCCs and RCCs did not perform as well as MFCCs and this is probably due to the fact that the speech coding in mobile phone networks removes relevant information about the glottal shaping and lip radiation components of the speech signal. These components are

essential to the performance of these parameters and removing them could be expected to result in a worse FVC performance.

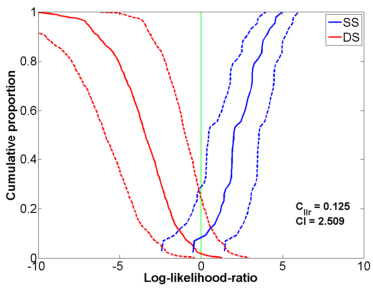

Figure 2: *Tippett plot showing the performance of CCCs for AMR coded speech at 12.2 kbps*

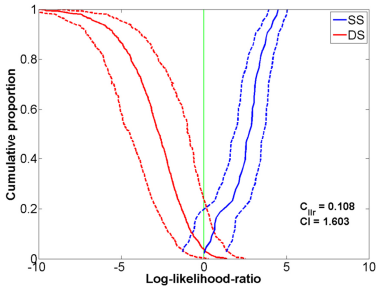

Figure 3: *Tippett plot showing the performance of MFCCs for AMR coded speech at 12.2 kbps.*

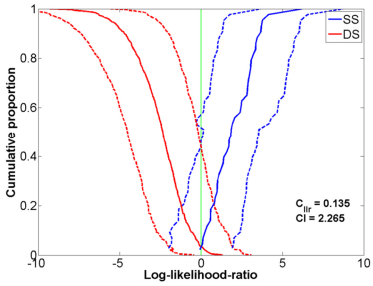

Figure 4: *Tippett plot showing the performance of CCCs for EVRC coded speech using OP0.*

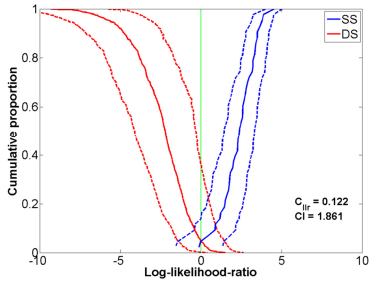

Figure 5: *Tippett plot showing the performance of MFCCs for EVRC coded speech using OP0.*

However, CCCs were still the second best performing parameter followed by RCCs and then LPCCs. The performance of formant trajectories was the worst amongst all the parameters. Automatic parameters are known to work well in speaker recognition because, being estimated over a sufficiently long portion of speech, they are able to average out linguistic information. Though we have used shorter segments of speech, this approach has still worked well, which is unexpected and requires further investigation.

7. References

[1] Alzqhoul, E.A., B.B. Nair, and B.J. Guillemin, *Speech Handling Mechanisms of Mobile Phone Networks and Their Potential Impact on Forensic Voice Analysis*, in *SST 2012, Sydney, Australia,* 2012.

[2] Guillemin, B.J. and C. Watson, *Impact of the GSM Mobile Phone Network on the Speech Signal–Some Preliminary Findings.* International Journal of Speech Language and the Law, 2008. **15**(2): p. 193-218.

[3] 3GPP, *TS 26.071 V11.0 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech CODEC speech processing functions; AMR speech CODEC; General description Retrieved on 2 June 2013, last retrieved from http://www.3gpp.org/.* 2012a.

[4] 3GPP2-EVRC, *E.V.R. Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems General description Retrieved on 2 June 2013, last retrieved from* http://www.3gpp2.org/, 2007.

[5] Nair, B.B., E.A. Alzqhoul, and B.J. Guillemin, *Determination of likelihood ratios for forensic voice comparison using principal component analysis*, International Journal of Speech Language and the Law, vol. 21(1), pp 83-112, 2014.

[6] Morrison, G.S., *Forensic voice comparison.* Expert Evidence, 2010. **40**: p. 1-105.

[7] Morrison, G.S., *Measuring the validity and reliability of forensic likelihood-ratio systems.* Science & Justice, 2011. **51**(3): p. 91-98.

[8] Morrison, G.S., T. Thiruvaran, and J. Epps. *Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system*. in *Proceedings of Odyssey*. 2010.

[9] 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.

[10] Gonzalez-Rodriguez, J., et al., *Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition.* Audio, Speech, and Language Processing, IEEE Transactions on, 2007. **15**(7): p. 2104-2115.

[11] 3GPP, *TS 45.009 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network;Link adaptation.Retrieved on 20 June 2013, last retrieved from http://www.3gpp.org/.* , 2012c.

[12] Rabiner, L.R. and R.W. Schafer, *Theory and application of digital speech processing*. International Edition. 2009: Pearson.

[13] Oppenheim, A. and R. Schafer, *Homomorphic analysis of speech.* Audio and Electroacoustics, IEEE Transactions on, 1968. **16**(2): p. 221-226.

[14] Papamichalis, P.E., *Practical approaches to speech coding*. 1987: Prentice-Hall, Inc.

[15] Fant, G., *Acoustic theory of speech production*. 1970: Walter de Gruyter.

[16] Snell, R.C. and F. Milinazzo, *Formant location from LPC analysis data.* Speech and Audio Processing, IEEE Transactions on, 1993. **1**(2): p. 129-134.

[17] Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999. Citeseer.

[18] Ramos-Castro, D., J. Gonzalez-Rodriguez, and J. Ortega-Garcia. *Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework*. in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. 2006. IEEE.