Impact of Various GSM Network Factors on Forensic Voice Comparison

Balamurali B. T. Nair [1,2], *Esam A. S. Alzqhoul* [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
bbah005@aucklanduni.ac.nz, ealz002@aucklanduni.ac.nz, bj.guillemin@auckland.ac.nz

Abstract

Speech transmitted through the GSM network can be negatively impacted by a number of factors, such as Dynamic Rate Coding (DRC), Frame Loss (FL) and Background Noise (BN) at the transmitting end. This paper reports on a study to investigate which of these has the greatest impact on the results of a Forensic Voice Comparison (FVC). It is shown that FL tends to have the most significant impact.

Index Terms: Forensic Voice Comparison (FVC), GSM network, Dynamic Rate Coding (DRC), Frame Loss (FL) and Background Noise (BN)

1. Introduction

Mobile phones are now a widely used means of communication amongst the criminal fraternity. In the mobile phone arena there are a number of network systems used worldwide and amongst these the Global System for Mobile Communication (GSM) is currently the most popular with 4.4 billion subscribers [1]. The speech signal transmitted through the GSM network is negatively impacted by a number of factors, such as Dynamic Rate Coding (DRC) [2], Frame loss (FL) [3] and Background Noise (BN) at the transmitting end [4, 5]. Of these, the one that affects FVC the most is unclear and this is investigated in this paper.

To examine the impact of these factors, one can adopt one of two experimental strategies. The first involves transmitting speech through an actual network. This approach, however, has a major drawback in that it permits investigation for only a finite set of transmission conditions existent during a particular call or set of calls. In reality, channel conditions can vary significantly from one call to the next and from one location to the next. For the results of such an investigation to be meaningful, the totality of all possible transmission scenarios needs to be included, not a small subset. Furthermore, with this approach it is not possible to examine the impact of a particular aspect in isolation to others.

The second strategy is to focus on the speech codec implemented in these networks and drive a software implementation of it under all of its possible modes of operation. The rationale for this approach is that while it is the network which dynamically decides the operating mode according to such factors as changing channel quality, it is the codec which implements it. Thus the codec is solely responsible for the quality of the transmitted speech signal [6]. We consider this latter approach to more comprehensively reflect the impact of each and every aspect of the network on speech and for this reason it was chosen for this investigation.

The most widely used speech codec in the GSM network is the Adaptive Multi-Rate (AMR) codec. It operates on speech sampled at 8 kHz and codes it into 20 ms frames. The speech coding technique used is called Algebraic Code Excited Linear Prediction (ACELP) and it can code individual frames 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 [7].

In this investigation the strength of speech evidence was evaluated using the Bayesian likelihood ratio (LR) framework. The LR is a measure of the probability of the evidence (i.e., the suspect and offender data) given the competing same-origin (prosecution) and different-origin (defense) hypotheses [8, 9]. Principal Component Analysis Kernel Likelihood Ratio (PCAKLR) was used to calculate LR values [10].

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in automatic speaker recognition and FVC for speech acquired from a variety of sources. They have been shown to be optimum when analyzing mobile phone speech as well [11], so were used for this investigation. The accuracy of a FVC analysis has been estimated using the Cost Log-likelihood Ratio (C_{llr}) and its reliability using Credible Interval (CI) [12]. Results have also been shown graphically using Tippett plots [13].

The remainder of this paper is structured as follows. An overview of the factors in a GSM network that can impact a FVC analysis is presented in Section 2. Section 3 discusses the experimental methodology chosen for this investigation. Results and findings are presented in Section 4, followed by conclusions in Section 5.

2. Overview of various GSM factors

2.1. Dynamic rate coding (DRC)

DRC is the process of changing the speech coding bit rate dynamically in accordance with changing channel conditions, and, to a lesser extent, changing channel congestion (i.e., number of users). The coding bit rate in turn directly impacts the quality of the resulting speech signal, and thus any subsequent FVC analysis. With the goal of maintaining a certain minimum speech quality irrespective of changing channel conditions, the GSM network instructs the codec to adjust its coding bit rate either up or down as necessary [2].

DRC is implemented in two stages: channel mode adaptation followed by codec mode adaptation. The first is determined by the number of users (i.e., congestion) in a cell. Full rate (TCH/FR) mode is selected when congestion is low and half rate (TCH/HR) when it is high. With TCH/FR any of the codec's eight bit rates can be used; with TCH/HR this is restricted to the lowest five bit rates. How frequently bit rate can be changed is dependent on which codec mode adaptation procedure is in force. With ETSI-specified fast link adaptation, this is a maximum of every 40 ms, while with Nokia proprietary slow link adaptation the maximum is every 480 ms [14, 15]. A detailed analysis of the impact of DRC alone on FVC can be found in [2].

2.2. Frame loss (FL)

The wireless channel in a mobile network is often very poor, increasing the likelihood of speech frames being either lost or irrecoverably corrupted during transmission. With the GSM network no distinction is made between these two. Whenever a frame is corrupted, an attempt is made to correct it using error correction strategies. If this fails, or the frame is lost, it is synthetically replaced at the receiving end using a history of past ‘good’ frames [16]. When a long sequence of frames is lost, the amplitudes of replaced frames are gradually decreased until silence results or the call is dropped. A detailed analysis of the impact of FL in isolation on FVC can be found in [3]

2.3. Background noise (BN)

BN is frequently present in mobile phone communications. It is different from channel noise in that it originates from a variety of sources at the caller’s location. When added to the speech signal, it negatively impacts the coding process, resulting in a degradation of perceptual speech quality. Though channel noise is also present, it can never directly impact the transmitted speech signal, but rather indirectly by causing frames to get corrupted or lost. As a consequence, all received speech frames are noise-free, though some will be noise-free synthetic replacements for frames either lost or too badly corrupted.

Unlike codecs used in some other networks, the AMR codec has no mechanism for mitigating the impact of BN on the coding process [17]. The severity of this will vary depending on the BN type and the resulting Signal-to-noise ratio (SNR). A detailed analysis of the impact of BN in isolation on FVC can be found in [4].

3. Experiment Setup

3.1. Speech database

The XM2VTS database containing speech recordings of 295 subjects (156 male, 139 female) was used for this investigation [18]. The language is English with predominantly a Southern British accent. Subjects were recorded on four different occasions separated by one month intervals. During each session each subject repeated three “sentences” twice. The first two sentences 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 sentence was: “Joe took Father’s green shoe bench out”.

Of the 156 male speakers, 26 were discarded as either their recordings sounded less audible or they were judged to have a different accent from the rest. The remaining 130 speakers were divided into three sets: 44 speakers in the Background set, and 43 speakers in each of the Testing and Development sets. This resulted in 43 same-speaker (SS) comparisons and 903 different-speaker (DS) comparisons. Further, three different recording sessions permitted two SS and three DS comparisons for every speaker, from which CI values were determined.

Speech files were down sampled to 8 kHz to align with the input requirements of the codec and the three words – “nine”, “eight” and “three” – were extracted from each speaker’s first three recording sessions. Using a combination of auditory and acoustic analysis, the vowels /ai/, /ei/ and /i/ were then extracted. Including diphthongs and a monophthong in the investigation was important because codecs in mobile phone networks often code stationary and non-stationary speech segments differently. A single feature set comprising 23 MFCCs was determined from the entire duration of each vowel segment and used as input to PCAKLR to produce a score. Scores were then calibrated and fused using logistic regression to produce LR values, the required calibration and fusion parameters being determined from the Development set comparisons [19]. A mean LR was calculated for every comparison (i.e., the mean of two LRs for SS comparisons and the mean of three LRs for DS comparisons). These mean values were then used for computing C_{llr}.

3.2. Experimental Methodology

Three FVC experiments were undertaken, as shown in Figure 1. Experiment 1 (top branch) used un-coded speech; Experiments 2 and 3 (middle and lower branches, respectively) used AMR-coded speech. The difference between Experiments 2 and 3 related to which mobile phone factors were incorporated (DRC, FL and BN for Experiment 2; DRC and BN for Experiment 3). In each experiment the Background set was identically processed to the speech being compared.

3.2.1. Implementing DRC

In respect to DRC, a medium-channel-quality scenario has been chosen. With this the codec can use any of its eight bit rates. Which ones it actually uses for a particular call is determined by a complicated process (see [2] for a detailed discussion). But in brief, it involves selecting only four of its eight bit rates, this set being referred to as the Active Codec Set (ACS). With eight bit rates, the total number of possible ACS combinations is 162. One of these combinations was randomly chosen for coding a particular vowel token and it was then dynamically coded using the corresponding four bit rates. The medium-channel-quality scenario also includes a constraint on the initial bit rate to use in a particular call as well as constraints associated with switching to the nearest neighbours in a selected ACS. All these constraints have been included in this investigation.

The ETSI-specified fast link adaptation, permitting bit rate changes a maximum of every 40 ms, has been used. The Nokia proprietary slow LA has not been investigated here as it can be considered to be a subset of the ETSI-specified LA scheme where the chosen codec set contains only one bit rate [2].

3.2.2. Implementing FL

In respect to FL, a frame error rate (FER) in the region of 10 to 15%, which approximately translates into a Mean Opinion Score (MOS) of 2.9 [3], has been used. This equates to the lowest voice quality permitted in such networks. When voice quality drops below this, a call is terminated automatically. Given that the durations of the vowel segments used in these experiments were of the order of 12 to 15 frames, this FER translates into a maximum number of lost frames per vowel segment being typically one, or at most two. Worst-case conditions have been chosen (i.e., two lost frames per vowel token), their locations being determined randomly according to a uniform distribution [3].

3.2.3. Adding BN

In respect to BN, the designers of mobile phone networks typically undertake performance tests using three types of noise: car, babble and street noise, and at three SNRs: 9, 15 and 21 dB [4]. A recent investigation has found that babble

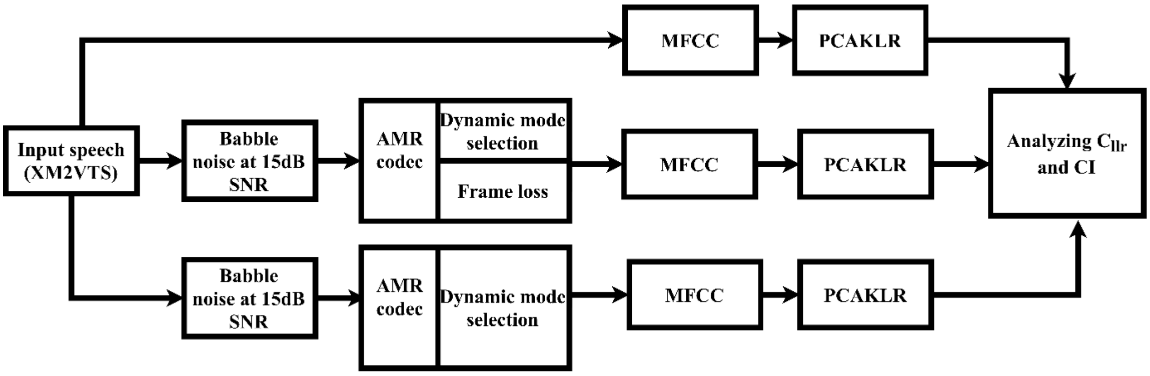

Figure 1: *Block diagram of the experimental setup*

noise tends to have a higher impact on FVC in a GSM network than the other two [4], and so babble noise was used in this investigation. The SNR of the speech signal was set to a medium value of 15 dB.

4. Results

Table 1 shows results of these experiments. Considering first the mean C_{llr} values and comparing performance between un-coded and coded speech, it is evident that AMR coding adversely affects the accuracy of a FVC when DRC, FL and BN are present (compare Experiment 1 with 2 – a lower C_{llr} value translates to better accuracy). However, when DRC and BN are present in the coded speech, but FL is removed, FVC accuracy has very slightly improved compared to the un-coded case (compare Experiment 1 with 3).

Table 1. *Impact of various factors of the GSM network on FVC performance.*

Speech Condition	Mean C_{llr}	CI
Exp. 1: Un-coded	0.167	2.299
Exp. 2: AMR-coded with DRC, FL & BN	0.216	1.627
Exp. 3: AMR coded with DRC & BN (FL excluded)	0.166	1.772

Firstly addressing the slight improvement in FVC accuracy between un-coded and coded speech with FL removed, this is clearly unexpected, but is a result we have observed many times when working with mobile phone speech [2, 3, 4]. We conjecture that it is related to the quantization processes inherent in the AMR-coding which tend to remove small variations in the speech parameters and thereby reduce the differences between voices samples. It seems reasonable to expect this to impact SS comparisons more than DS comparisons, making them even more similar, though both will of course be affected. The fact that, notwithstanding this apparent beneficial impact of the codec, accuracy is quite a bit worse when FL is included in the coded speech, suggests that the impact of FL is more significant than for either DRC or BN.

Focusing now on the CI values in Table 1, it is clear that the reliability of a FVC is better for AMR-coded speech compared to un-coded (the lower the CI value, the better the reliability). We again conjecture that this is due the same quantization processes mentioned above and for the same reasons. But there is little difference in CI for coded speech when FL is present or it is excluded.

Some insight into the above observations can be obtained by examining the Tippett plots for these experiments (Figures 2, 3 and 4 corresponding to Experiments 1, 2 and 3, respectively). These show the cumulative distributions of LLR values for SS (solid blue curve) and DS (solid red curve) comparisons, where $\mathrm{LLR} = 10\mathrm{Log}_{10}\mathrm{LR}$. (Note: LLRs for correctly classified SS comparisons are positive, those for DS are negative. The greater their magnitude, the stronger the evidence either way). The corresponding dotted lines show the 95% confidence interval for the LLRs.

Comparing first Figure 2 for un-coded speech with Figures 3 and 4 for coded speech, it is clear that the magnitudes of the LLRs for the un-coded case tend to be higher than for the coded, from which it can be concluded that AMR-coding generally negatively impacts the strength of the evidence for a FVC, which is an expected result. Whether this is in part due to the quantization processes mentioned previously is an aspect which deserves further investigation. These plots also confirm that reliability is somewhat better for coded speech than for un-coded.

Comparing now Figures 2 and 3, the proportions of contrary-to-fact LLRs for both SS and DS comparisons are higher for coded speech when FL has been included. Further, SS comparisons have been affected slightly more in this regard than DS comparisons. Comparing Figure 2 with Figure 4 shows that when FL is removed in the coded speech, the proportions of contrary-to-fact LLRs for both coded and un-coded speech are very similar. (Note: to align these observations with the C_{llr} values shown in Table 1, it needs to be remembered that the C_{llr} performance measure gives a greater weighting to LLRs closer to the LLR=0 boundary than those further away, irrespective of whether they are consistent-to-fact or contrary-to-fact.)

5. Conclusions

There are three major factors that impact speech transmitted through a mobile phone network: Dynamic Rate Coding, Frame loss and Background Noise at the transmitting end. This paper has reported on an investigation to determine which of these associated with the GSM mobile phone network has the greatest impact on the performance of a FVC analysis in terms of accuracy and reliability. It has been shown that Frame Loss has the greatest impact on the accuracy of such an analysis and that this seems to be linked to a greater proportion of same-speaker classifications which are contrary-to-fact than different-speaker. In terms of reliability, the coding process associated with this network actually seems to improve this aspect, an observation we conjecture is linked to the associated quantization processes involved.

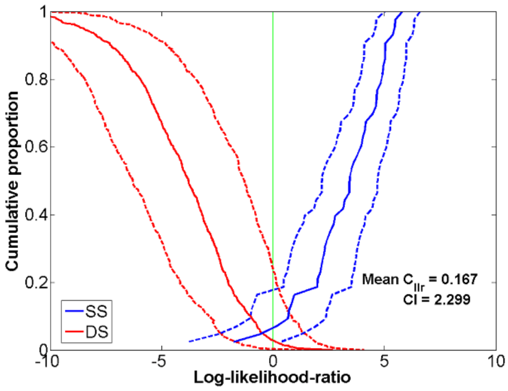

Figure 2: *Tippett plot of the performance for un-coded speech.*

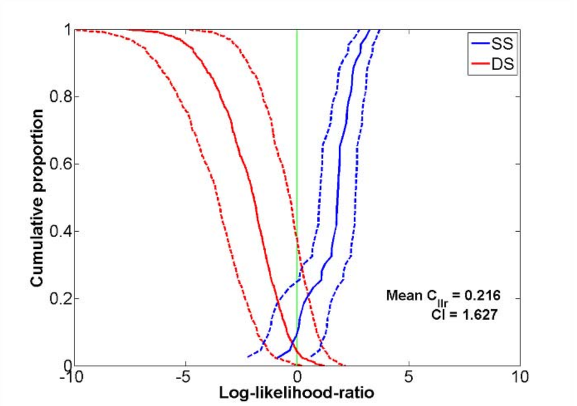

Figure 3: *Tippett plot of FVC performance for coded speech incorporating DRC, FL and BN.*

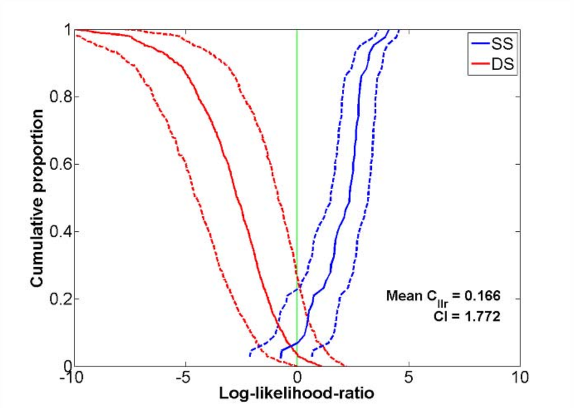

Figure 4: *Tippett plot of FVC performance for coded speech incorporating DRC and BN, but with FL excluded.*

6. References

[1] www.gsacom.com, "GSM celebrates 20 years," *Retrieved on 2 May 2016, last retrieved from http://networks.nokia.com/news-events/press-room/press-releases/gsm-celebrates-20-years,* 2011.

[2] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison," *Science & Justice,* vol. 55, pp. 363-374, 2015.

[3] B. B. Nair, E. A. Alzqhoul, and B. J. Guillemin, "Impact of frame loss aspects of mobile phone networks on forensic voice comparison," *International Journal of Sensor Networks and Data Communications,* Vol. 2015, 2015.

[4] B. B. Nair, E. A. Alzqhoul, B. J. Guillemin "Impact of background noise in mobile phone networks on forensic voice comparison," *J Forensic Leg Investig Sci,* Vol. 2: 007, 2016.

[5] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Speech handling mechanisms of mobile phone networks and their potential impact on forensic voice Analysis," presented at *Australasian Speech Science and Technology* Conference, 2012, Sydney, Australia, 2012.

[6] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "An alternative approach for investigating the impact of mobile phone technology on speech," in *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*, 2014.

[7] 3GPP, "TS 26.101 V11.0.0 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech codec speech processing functions;Adaptive Multi-Rate (AMR) speech codec frame structure. Retrieved on 2 May 2013, last retrieved from http://www.3gpp.org/," ed, 2011.

[8] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* Vol. 15, pp. 2104-2115, 2007.

[9] G. S. Morrison, "Forensic voice comparison," *Expert Evidence,* Vol. 40, pp. 1-105, 2010.

[10] B. Nair, E. Alzqhoul, and B. J. Guillemin, "Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis," *International Journal of Speech, Language & the Law,* Vol. 21, 2014.

[11] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Comparison between speech parameters for forensic voice comparison using mobile phone speech," *Proceedings of the Australasian Speech Science and Technology Association, Christchurch,* pp. 29-32, 2014.

[12] G. S. Morrison, "Measuring the validity and reliability of forensic likelihood-ratio systems," *Science & Justice,* Vol. 51, pp. 91-98, 2011.

[13] D. Meuwly 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.

[14] Nokia, "Guidelines for practical implementation of AMR in Nokia's Network element. General description. Retrieved on 21 June 2013, last retrieved from http://www.scribd.com/doc/104330368/14/Initial-codec-mode-selection," ed, 2004.

[15] 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/. ," ed, 2012c.

[16] 3GPP, "TS 26.091 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Error concealment of lost frames. Retrieved on 6 April 2013, last retrieved from http://www.3gpp.org/," ed, 2012b.

[17] 3GPP, "3GPP TS 26.077, Minimum performance requirements for noise suppresser; Application to the Adaptive Multi-Rate (AMR) speech encoder Retrieved on 2 June 2013, last retrieved from http://www.3gpp.org/," 2012.

[18] J. Lüttin, "Speaker verification experiments on the XM2VTS database," in *IDIAP-RR 99-02, IDIAP*, 1999.

[19] D. Ramos-Castro, 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, pp. 1-8.