Journal of Forensic Research
ISSN: 2157-7145

Nair BBT et al, J Forensic Res 2016, 7:2
DOI: 10.4172/2157-7145.1000324

Research Article Open Access

Impact of the GSM and CDMA Mobile Phone Networks on the Strength of Speech Evidence in Forensic Voice Comparison

Balamurali B T Nair[1,2]*, **Esam A S Alzqhoul**[1,2] **and 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*

Corresponding author: Balamurali B T Nair, Ph D., The University of Auckland, New Zealand, E-mail: bbah005@aucklanduni.ac.nz

Received date: Nov 22, 2015;**Accepted date:** May 11, 2016; **Published date:** May 25, 2016

Abstract

It is becoming increasingly common for mobile phone recordings to be presented as evidence in a court of law. In such situations a forensic scientist is frequently engaged to analyse suspect and offender voice samples with a view to determining the strength-of-evidence, a process called Forensic Voice Comparison (FVC). This paper investigates the extent to which an FVC analysis is negatively impacted by the network through which the speech has passed. Our investigation focuses on the GSM and CDMA networks as these are the ones in common usage currently. Our experimental findings suggest that both networks negatively impact the accuracy of an FVC analysis and that this is worse for the CDMA network. We present strategies for mitigating this impact to some extent.

Keywords: GSM; CDMA; AMR; EVRC; Forensic voice comparison; Likelihood ratio

Introduction

There is no doubt that non-registered mobile phones are becoming a popular choice for criminals to conduct their anonymous calls and intercepting these calls can help police officers bring criminals to justice. In such cases forensics speech scientists are often engaged to assist in revealing the offender identity by examining the similarities and differences between suspect and offender mobile phone speech recordings, a process referred to as Forensic Voice Comparison (FVC) [1]. The forensic scientist might, however, undertake their analysis under the erroneous assumption that all mobile phone networks impact the speech signal in a similar manner. There are two major technologies currently in use in the mobile phone arena: Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA). These networks are very different in their ways of handling, processing and transmitting speech and therefore in their respective impacts on the speech signal [2]. The primary goal of this paper is to examine the ways and extent to which these two technologies can impact the strength of speech evidence associated with a FVC analysis. Knowledge of these impacts is of critical importance to practitioners in this arena when presenting their findings to a court.

The key aspects in these networks that can directly or indirectly impact the speech signal are: (i) Dynamic Rate Coding (DRC), (ii) Frame Loss or corruption (FL), and (iii) Background Noise (BN) at the transmitting end [2-5]. Though all three are clearly of concern in this arena, FL and BN are particularly so. Before discussing the specifics of each of these aspects, it is important to note that the speech codec in these networks is the only component that is directly responsible for handling the speech signal and thus any changes that might occur to it. These codecs have many modes of operation, but changing these modes is a process initiated by the network as a whole in response to changing channel/capacity conditions and changing speech characteristics. The most widely used speech codecs in the GSM and CDMA networks are the Adaptive Multi Rate (AMR) codec [3] and the Enhanced Variable Rate Codec (EVRC) [6], respectively.

DRC [7-9] is the process of changing the source coding bit rate for each 20 ms frame in accordance with changing channel conditions (i.e., quality) in the case of the GSM network, or increases in the number of users accessing the system (i.e., capacity) in the case of the CDMA network. When wireless channel conditions are bad, the GSM network instructs the AMR codec to increase the number of bits used for protection, while reducing those for coding the speech signal, thus resulting in a lower speech quality, the goal being to maintain a fixed data load on the network per call. When channel conditions improve, this strategy reverses. In the CDMA network, if a large number of users are sharing a cell site, the network instructs the EVRC codec to achieve a low Average Data Rate (ADR) over a certain number of speech frames, resulting in a relatively low speech quality. If the number of users decreases, the codec is permitted to operate at a higher ADR, resulting in a higher speech quality. Unlike the GSM network, though, in the CDMA network the decision about the actual source coding bit rate to be used per frame is left to the codec. The codec does this by analyzing the speech signal and classifying each frame into one of the following categories: voiced, voiceless (often referred to as unvoiced), transient or silence [6].

The second key aspect in a mobile phone network that directly impacts the speech signal is FL. In contrast to landline networks, the medium of transmission in mobile phone networks is wireless and is thus subject to many adverse factors which can result in frames being either lost or irrecoverably corrupted during transmission. Both eventualities are treated the same, there being two possibilities in respect to the corrective action taken. Either the codec replaces the lost or corrupted frame with the last good speech frame, or it generates a new one by means of extrapolation using a history of previous good frames [8,10]. If a long sequence of frames is lost, the amplitude of the replaced frames decreases until silence results or the call is terminated. The codec can insert up to 16 speech frames in this manner, which means that multiple segments of artificial speech of up to 320 ms

duration could be present in the recovered speech signal [11,12]. The only difference between the GSM and CDMA networks in this regard is that extrapolation is not used in the latter.

The third key aspect, namely BN, is separate and entirely unrelated to channel noise. Channel noise impacts the transmission process and typically results in FL [13,14]. BN originates at the transmitting end and thus mixes with the speech signal at that point. The two networks, and specifically their respective codecs, differ in how they handle BN. Unlike the EVRC codec, the AMR codec has no mechanism for countering the effects of BN or distinguishing BN from speech. As a result, a combination of the two is coded and transmitted in the GSM network. With the EVRC codec, however, a preprocessing stage called Noise Suppression (NS) is implemented to filter out BN from the input speech signal. When noise levels are not too high, this NS process can actually improve the perceptual quality of the transmitted speech signal. But in instances of high BN levels, such as would be the case when a call originates from a moving vehicle, it is less able to distinguish between speech and noise, typically resulting in removal of sections of the original speech signal [15]. Therefore, it can be anticipated that FVC performance will be generally lower for the CDMA network under high levels of BN.

To investigate the impacts of mobile phone networks on speech, two approaches are possible. The first of these involves transmitting speech through an actual network and then analyzing the received waveform. But this would need to be done a very large number of times in order to capture a representative sample of all possible transmission scenarios that could take place during a call. Even then there would be no way of knowing whether this goal had indeed been achieved because the received speech signal carries no information of the actual transmission conditions present at any instant in time. The second approach involves passing the speech signal through a software implementation of the speech codec in a highly controlled manner intended to be representative of the broad spectrum of all possible channel and BN conditions that could be present in an actual network. From the standpoint of a robust scientific investigation, we regard the latter approach to be superior, so is the one adopted for this investigation.

One approach to conducting a FVC analysis is using the likelihood ratio (LR) framework [1,16-18]. Within this framework different methods have been proposed to evaluate the speech evidence, such as Multivariate Kernel Density (MVKD) [19], Gaussian Mixture Model-Universal Background Model (GMM-UBM) [20], and Principle Component Analysis Kernel Likelihood Ratio (PCAKLR) [21,51]. These methods incorporate probabilistic models comprising two major terms: similarity and typicality. The former quantifies the similarity between suspect and offender speech samples, the latter their typicality in respect to a relevant background population. MVKD and PCAKLR are primarily designed for token-based analysis, whereas GMM-UBM is primarily designed for data-stream-based analysis. In this investigation we have undertaken token-based experiments using 23 Mel-frequency cepstral coefficients (MFCCs). Given that MVKD is designed to work with a small number of parameters, typically 3 or 4, whereas PCAKLR can handle a much larger number of parameters, we have chosen to use PCAKLR in our experiments.

It is known that cepstral coefficients are generally sensitive to transmission artifacts in the landline phone network and several compensation techniques have been proposed to account for this [22-24]. Though transmission artifacts also impact the speech signal in mobile phone networks, as explained above, the manner in which they do so is completely different. Thus these compensation techniques might not be appropriate when working with mobile phone speech. The rest of this paper is structured as follows. A brief overview of both the LR framework and PCAKLR is presented in the following section. This section also describes the performance measuring tools used in our experiments to investigate the overall impact of mobile phone networks on the outcome of a FVC analysis. These tools are the loglikelihood-ratio cost (C_{llr}), Tippett plots, Applied Probability Error plots (APE) and Credible Interval (CI). This is followed by a discussion of experimental methodology. Finally, results and findings are presented and are followed with some concluding remarks.

Investigating the performance of a FVC analysis

Likelihood Ratio (LR) and PCAKLR

The evaluation of speech forensic evidence using the LR framework is gaining greater acceptance among forensic speech scientists [1,16,18,25-27]. This framework provides a quantitative answer to the following question: How much more likely is it to observe the properties of the offender and suspect speech samples assuming they have the same origin (prosecution hypothesis) than a different origin (defense hypothesis).

Mathematically, the LR is the ratio of two conditional probabilities:

$$LR=\frac{p(E|Hp)}{p(E|Hd)}$$

where E is the evidence and p(E|Hp) and p(E|Hd) are the conditional probabilities of the evidence given the prosecution and defense hypothesis, respectively. LR values significantly greater than one support the prosecution hypothesis; values significantly less than one support the defense hypothesis; values close to one provide little support either way. Log–likelihood ratios (LLRs) are usually calculated from LRs, where LLR is computed as $\log_{10}$ (LR). The magnitude of the LLR is a measure of the strength-of-evidence and its sign indicates whether this in the favor of the prosecution hypothesis (positive values) or defense (negative values).

PCAKLR is a two stage process. Firstly a set of input parameters is transformed into another set of orthogonal (i.e., highly uncorrelated) parameters (called transformed parameters) using principal component analysis (PCA) [28-30]. Next an LR value for each of these transformed parameters is determined using univariate kernel density (UKD) analysis [31] their product being taken to produce an overall LR value.

Log-likelihood-ratio cost (C_{llr})

C_{llr} is one of the metrics widely used in speech forensics to measure the validity (or accuracy) of a FVC analysis [1,18,26,32]. This is estimated by calculating the LR output for a large set of test-sample pairs for which same-speaker and different-speaker origins are known. These results are then compared to the actual outputs to reveal the accuracy of a FVC system. Mathematically, it can be calculated as:

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}}\log_2\left(1+\frac{1}{LR_{soi}}\right)+\frac{1}{N_{do}}\sum_{i=1}^{N_{do}}\log_2\left(1+LR_{doi}\right)\right)$$

where N_{so}, N_{do} are the number of same- and different-speaker comparisons, respectively, and LR_{so}, LR_{do} are the LRs determined for

same- and different-speaker origins, respectively. The lower the C_{llr} value, the more accurate is the analysis, and vice versa.

Tippett plots

The LLR results of a FVC analysis can be graphically represented using Tippett plots [1,33]. With these (examples shown in results), the solid blue line rising towards the right shows the cumulative proportion of same-speaker comparisons which have LLR values greater than or equal to the corresponding value on the horizontal axis.

The solid red curve rising towards the left indicates the cumulative proportion of different-speaker comparisons with LLR values less than or equal to the value indicated on the horizontal axis. Since large positive LLRs support the same-speaker hypothesis and large negative LLRs the different-speaker hypothesis, the further apart these curves are, the better the results.

Applied Probability of Error (APE) plots

Losses in the accuracy of a FVC analysis can be investigated using APE plots [34,35]. If the system under evaluation is lossless, then C_{llr} = 0 would result. In reality all systems do have losses and these consist of two parts: discrimination loss (C_{llrmin}) and calibration loss (C_{llrcal}). C_{llrmin} is the lowest C_{llr} that could be achieved by optimizing the LR values under evaluation. C_{llrcal} is the difference between the actual C_{llr} and C_{llrmin}.

An APE plot comprises a set of APE curves and bar graphs. The APE curves plot the error rate against logit priors. The height of different shades in the bar graph represents the area under the APE curves.

Three different kinds of APE curves are plotted. The green curve corresponds to the error-rate of the optimized LLRs and the area under this curve is reflected in the height of the green bar graph.

The red curve refers to the error rate of the actual LLRs. The area between the red and the green curves is shown as the height of the red bar graph. Finally the dashed curve refers to the error rate of a reference system (i.e., LLRs = 0) and this does not have a corresponding bar graph in the APE plot.

Credible Interval (CI)

Reliability (or precision) quantifies the amount of variation in a LR calculation [32]. The CI proposed in [36,37] has been used in our experiments to measure this. The CI calculation attempts to answer the question: how much variability in estimating the strength-of-evidence can be expected due to variability in the measurement of speech parameters if the comparison process was repeated several times across different recording sessions?

Two approaches can be used to estimate the CI: parametric and non-parametric. The parametric approach is used if homoscedasticity (i.e., variation in LR results is the same for all comparisons) can be assumed.

However, heteroscedasticity is typically the case with PCAKLR and thus the non-parametric CI calculation has been used in our experiments. It is usual practice to represent CI as a function of LLR using a series of dotted lines on the Tippett plot either side of the solid same- and different-speaker curves.

Experimental Methodology

Speech database and speech parameters

The XM2VTS database [38], which contains speech recordings of 297 speakers (both male and female), has been used in our experiments. We have chosen to use male speakers in our investigation. Of the 156 male speakers in the database, 26 were discarded as they sounded less audible or appeared to have a different accent (the language of the database is English with a Southern-British accent). Speakers in this database were recorded on four different occasions separated by one month intervals. During each session each speaker spoke two sets of numbers as well as a sentence: "zero one two three four five six seven eight nine", "five zero six nine two eight one three seven four" and "Joe took father's green shoe bench out". The speech files are sampled at 32 kHz and 16 bits digitized.

In our experiments the three words "nine", "eight" and "three" were extracted using audio editing software such as Goldwave [39] and Wavesurfer [40]. The corresponding vowel tokens /aI/, /eI/ and /i/ were then extracted from these using auditory and acoustic procedures [17]. These vowel tokens were lastly down sampled to 8 kHz and stored into 16 bits to align with the codec' input requirements [41,42].

Though this database contains four different recording sessions, only three of these have been used in our experiments. In summary, four tokens of three different vowels (two diphthongs and one monophthong) from 3 non-contemporaneous recordings have been used in our experiments. The 130 speakers were divided into three groups: 44 speakers in the Background set, 43 speakers in the Development set and 43 speakers in the Testing set. With 43 speakers in the Testing set, 43 same-speaker comparisons and 903 different-speaker comparisons are possible. Two same-speaker comparison results were obtained for each speaker in the Testing set by comparing their Session 1 recording with their own recordings in Sessions 2 and 3.

Three different-speaker comparisons were produced for each speaker by comparing their Session 1 recording with all other speakers' recordings from Sessions 1, 2 and 3. The Background set remained the same across all comparisons and included two recording sessions for all 44 speakers. The Development set was used to train the logistic regression-fusion system [43], the resulting weights of which were used to combine LRs determined from the Testing set. MFCCs were computed from the vowel tokens. MFCCs are linked to the perceptual aspects of speech and have been reported as the most appropriate parameters for use in FVC when dealing with mobile phone speech [44]. The extraction of MFCCs was as follows. Firstly a Hamming window was applied to the whole vowel segment to remove edge effects, followed by taking the Discrete Fourier Transform (DFT).

The resulting spectrum was segmented into bands and a set of triangular weighting functions (Mel-filter banks) were then applied. The amount of energy existing in each filter bank was estimated, and then a Discrete Cosine Transform (DCT) was applied to the logarithm of these energies, resulting in a set of 23 MFCCs [45]. It is quite common in FVC to use the first 12-14 MFCCs extracted from a stationary speech frame, along with their deltas and delta-deltas. However our previous experiments have shown that such MFCCs introduce a high variation in the resulting LRs. Conversely, a larger number of MFCCs when extracted from an entire vowel segment, even if it is non-stationary, have been shown to produce less variation, as well as give better FVC accuracy [46].

Experimental Procedure

Figure 1 shows a block diagram of our experimental procedure which involves a comparison between two sets of FVC analyses in terms of C_{llr} and CI. The first set is performed using speech that has not been processed by a mobile phone codec and referred to in this paper as clean or no-coded speech. The second set uses speech that has been processed either by the GSM or CDMA mobile phone codecs. The coding process has been performed incorporating DRC, FL and BN. Ideally one should conduct a large number of experiments incorporating all possible operational modes of the codec under various channel/capacity conditions. However, only typical and representative scenarios have been considered in this paper. Details of these experiments in respect to the GSM and CDMA networks are discussed next.

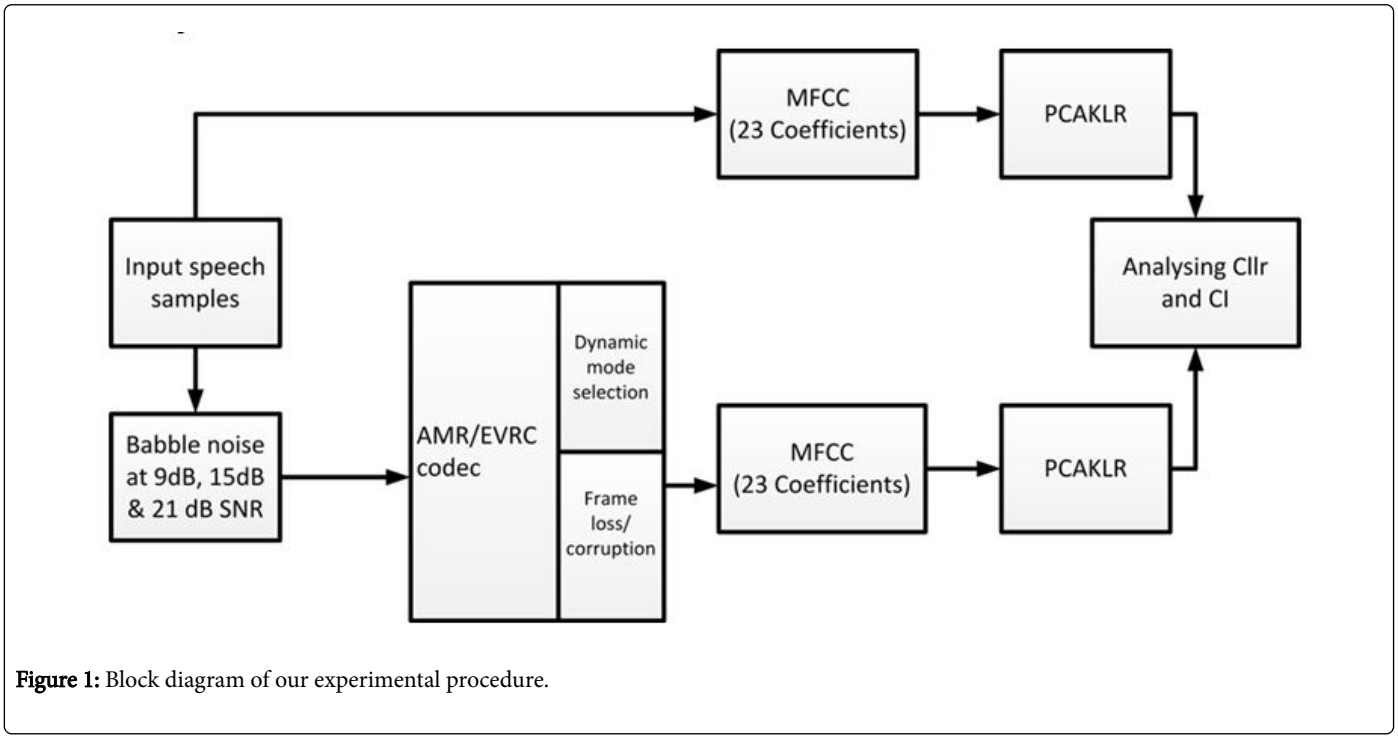

Figure 1: Block diagram of our experimental procedure.

For the experiment involving no-coded speech, all three data sets (i.e., Background, Development and Testing) used no-coded speech. But two different scenarios have been investigated in the case of coded speech, one without mismatch, the other with it. In the one without mismatch, all three data sets have used coded speech (i.e., no mismatch between Development/Testing and Background sets). In the one with mismatch, the Development and Testing sets have used coded speech, but the Background set has used no-coded speech (i.e., a mismatch exists between Development/Testing and Background sets). The goal of this latter experiment was to understand the extent to which such mismatch might impact the accuracy and reliability of an FVC analysis.

GSM-based experiments

For the GSM network speech files were coded using the AMR codec [3]. This codec can operate at eight bit rates (4.75, 5.15, 5.90, 6.70, 7.40 7.95, 10.20 and 12.20 kbps) and can be instructed by the network to switch between these every second speech frame (i.e., every 40 ms) [7,9]. In a particular call the switching between bit rates can only happen within a subset of a maximum four different bit rates, called the Active Codec Set (ACS). The ACS can be selected from either the lowest five bit rates for Half Rate (HR) channel mode, or from all the eight bit rates for Full Rate (FR) channel mode. The selection of these modes depends on the channel capacity (i.e., number of users present in a cell site). FR is selected when the number of users is small, HR when the cell site is congested with a large number of users [7].

When the channel condition is bad, low bit rates are selected for coding the speech signal. This can result in a relatively poor quality of speech, whereas when the channel condition improves the higher bit rates in the ACS are used resulting in better quality [7]. Therefore it can be assumed that in the case of medium channel conditions (i.e., neither good nor bad), the speech quality will be medium.

This scenario can be simulated by choosing the ACS from any of the eight bit rates and this is the speech quality used in our experiments. Thus, the incorporation of DRC into our experiments was achieved by selecting the bit rates for the ACS on a random basis from all eight bit rates and switching between bit rates within an ACS also randomly every second frame (i.e., every 40 ms), both processes conforming to a uniform distribution. In the absence of information to the contrary, the choice of a uniform distribution for both of these processes seemed reasonable.

In respect to incorporating FL into our GSM experiments, highly innovative techniques have been developed in these networks to try and mitigate the effects of this. When a speech frame is received, certain bits of the frame are checked to determine whether bit errors have occurred. If they have, the receiver uses convolutional coding in an attempt to correct them [47]. If unsuccessful, the frame is discarded and replaced with either the last good frame, or an artificially generated frame based on extrapolated data from previous good speech frames. The frequency of occurrence of this process is termed the Frame Error Rate (FER). If it exceeds 10-15%, it is known that the resulting voice quality is unpleasant to the listener and therefore the

call is dropped [11]. (Note: this same FER threshold is used in the CDMA network as well.) Given that our experiments used relatively short duration segments of speech (i.e., vowels) in the range of 12-15 frames, the aspect of FL was incorporated by introducing 2 lost frames per vowel segment, this being the worst-case scenario. The locations of these lost frames were determined randomly according to a uniform distribution, the assumption being that all frames are equally likely to get lost during transmission. As an aside, the FL mechanism is activated in the AMR codec by changing the value of specific flags in the decoder. The flag affected is "RX_TYPE" and this has been changed to "SPEECH_BAD", which indicates a damage to Class A bits in the received frame and such a frame must be treated as lost. This in turn sets another flag called BFI (Bad Frame Indicator) in the decoder to one [10].

The AMR codec includes no special provision for countering the negative impacts of BN, the third aspect incorporated into our experiments. Different kinds of BN are common in mobile phone calls, though some tend to have greater impact on the subsequent coding process than others [48]. The designers of mobile phone codecs typical test the performance of their products using three types of BN, namely car, street, and babble noise [13,14] and at three SNR levels: 9, 15, and 21 dB [49]. We have found babble noise to have the greatest impact on FVC performance when using mobile phone speech and hence this noise has been used in our experiments at the same SNR levels mentioned. Samples of babble noise were acquired from the Soundjay database [50].

CDMA-based experiments

The EVRC codec can produce speech frames at 0.8, 2, 4, and 8.55 kbps, the specific bit rate being in part determined by a frame's classification into voiced, voiceless, transient or silence [6]. The speech quality achieved is a function of the resulting Average Data Rate (ADR) which is decided by the network based on the number of users accessing the system (i.e., channel capacity). The ADR can take on any value between 4.8 kbps to 9.6 kbps. A target ADR is communicated to the codec every 20 ms. The codec is then left with the task of achieving this target using an appropriate mix of its available operational modes over successive frames. It has three modes, known as the Anchor Operating Points (AOPs): OP0, OP1 and OP2.

Appropriate switching between OP0 and OP1 produces a relatively high ADR and thus high quality speech; switching between OP1 and OP2 conversely produces a relatively low ADR and thus low quality speech. Under medium capacity conditions (i.e., the cell site is neither congested nor has only a few users), medium quality speech is produced as a result of switching between all three AOPs. This latter scenario is the one chosen in our experiments. It is appropriate for two reasons. Firstly, it aligns with our GSM-based experiments which focused on medium-quality speech. Secondly, OP1 and OP2 use a different set of coding algorithms than OP0 and we wanted to ensure that our experiments were as representative as possible of typical transmission scenarios. Discrete ADR values were used for coding the speech files to ensure a good coverage of the selected range. If a continuous range of ADR values had been used instead, it would increase the likelihood of those being concentrated more in one part of the range than another, resulting in a less uniform usage distribution of AOPs. Thus the incorporation of DRC into our EVRC-based experiments was achieved by randomly selecting target ADRs according to a uniform distribution from the group: 4.8, 5.8, 6.2, 6.6, 7.0, 7.5, 8.5, 9.6 kbps.

The incorporation of FL into our EVRC-based experiments was done in an identical manner to our AMR-based experiments (i.e., 2 lost frames per vowel segment, their locations being determined randomly according to a uniform distribution) and for the same reasons. As an aside, the FL mechanism in the EVRC codec is activated by changing the value of a specific flag "data_packet.PACKET_RATE" to "0xE" [6]. In respect to BN, this too was incorporated into our EVRC-based experiments in an identical manner to our AMR-based experiments (i.e., babble noise added to the speech files at 9, 15 and 21dB SNRs). In contrast to the AMR codec, however, the EVRC codec also includes Noise Suppression (NS) prior to the coding process, its aim being to subtract BN present in each input speech frame.

Results and Discussion

Table 1 compares the resulting FVC performance for a number of typical scenarios in the GSM and CDMA networks. The no-coded speech results are also shown. To facilitate a better comparison, a graphical presentation of these results is shown in Figures 2 and 3.

Network	**GSM**				**CDMA**			
Background set	**Matched**		**Mismatched**		**Matched**		**Mismatched**	
Cllr/CI	Cllr	CI	Cllr	CI	Cllr	CI	Cllr	CI
BN at 9dB SNR DRC, FL	0.209	1.988	0.300	1.224	0.279	1.681	0.506	0.855
BN at 15 dB SNR DRC, FL	0.216	1.627	0.263	1.877	0.249	2.157	0.358	1.396
BN at 21 dB SNR DRC, FL	0.173	1.727	0.187	1.681	0.181	1.648	0.343	1.201
No-coded speech	Cllr = 0.167 CI = 2.299							

Table 1: Comparative FVC performance between the GSM and CDMA networks.

Focusing first on the C_{llr} results, the following can be noted:

(i) the coding process has always resulted in a negative impact on the FVC accuracy compared to no-coded speech;

(ii) for both matched and mismatched conditions under all transmission scenarios, better FVC accuracy results for GSM-coded speech than CDMA-coded;

(iii) for both networks, C_{llr} improves as the SNR level increases, which is an expected trend;

(iv) for both networks, mismatched analysis always results in worse FVC accuracy than matched, with CDMA-coded speech being affected more in this regard than GSM-coded speech.

With respect to the reliability of LR results, the following can be noted:

(i) The coding process has actually improved this aspect, which is an unexpected result;

(ii) For both matched and mismatched conditions, CDMA-coded speech gives generally better FVC reliability than GSM-coded;

(iii) There appears to be no clear trend between SNR and FVC reliability;

(iv)Mismatched analysis generally results in better FVC reliability.

One important observation can be drawn from the results of Table 1.

From the standpoint of the forensic scientist, one would like a FVC analysis to be both accurate and reliable.

When working with mobile phone speech, one has no control over the network the recordings came from, or the transmission scenarios that occurred during the calls under analysis (i.e., DRC, FL and BN).

What one does have some control over, though, is the Background data set used in the analysis and specifically whether this is coded or no-coded.

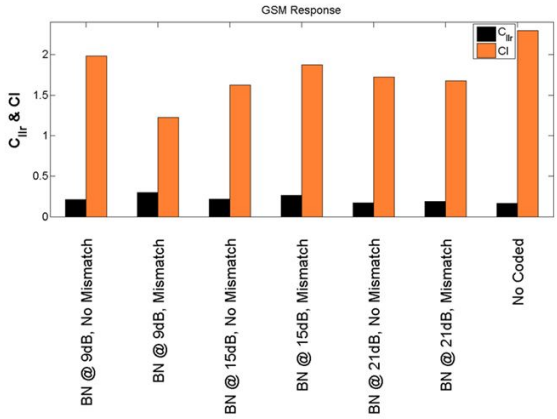

Figure 2: C_{llr} and CI results for the GSM network experiments.

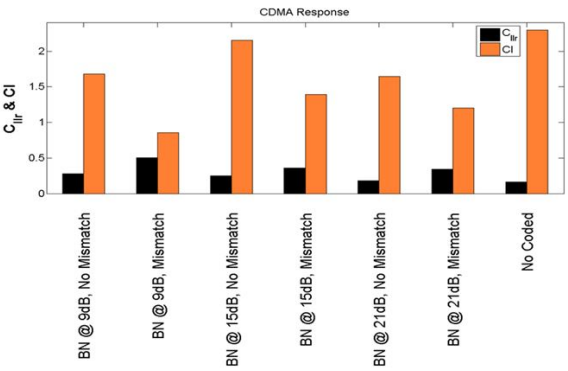

Figure 3: C_{llr} and CI results for the CDMA network experiments.

Assuming that FVC accuracy is more important than reliability, it would seem that using coded speech for the Background set is the better option to using no-coded. Returning now to the observation that coding appears to improve reliability, we conjecture that this may be related to the quantization processes inherent in the speech codecs. Quantization of speech parameters is likely to place some restriction on the range of values for a particular parameter, which may in turn lessen within-speaker variation slightly across different recording sessions while having comparatively little impact on between-speaker variation. This in turn may improve same-speaker comparisons. To further examine the FVC results, Tippett and APE plots have been produced for the investigated scenarios. Three representative Tippett plots are shown in Figures 4-6. The Tippett plot in Figure 4 shows FVC performance for the no-coded case. The GSM and CDMA coded speech results are shown in Figure 5 and Figure 6 respectively, for the case of SNR = 15 dB, 10-15% FL, and matched analysis. It is clear from Figures 4-6 that the proportions of different-speaker classifications which are contrary to fact are higher for coded speech than no-coded, and that CDMA-coded speech is worse in this regard than GSM-coded. This same trend was observed for the experiments conducted at other SNRs.

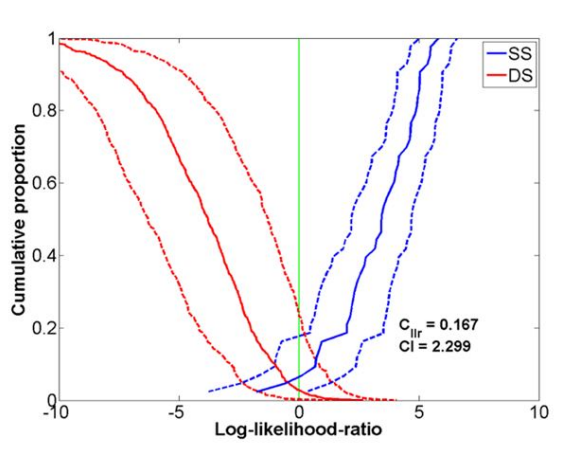

Figure 4: Tippett plot showing the performance of no-coded speech.

The Tippett plot in Figure 7 shows the LR results for CDMA-coded speech under mismatched conditions. It is interesting to compare this with the corresponding result of Figure 6 for matched conditions. Mismatched analysis has increased the proportions of same- and different-speaker classifications which are contrary to fact.

Further, it has had the effect of reducing the magnitude of the LRs overall (i.e., reducing the strength-of-evidence). Note the Tippett plot for GSM-coded speech under mismatched conditions is fairly similar to that of Figure 5 for matched conditions, so is not shown here. It reinforces the observation, though, that CDMA-coded speech is affected more by mismatch than GSM-coded.

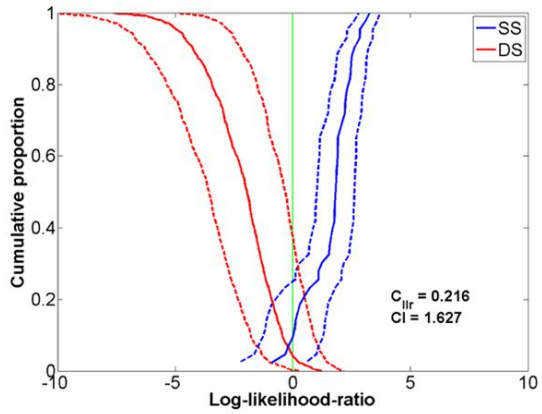

Figure 5: Tippett plot showing the FVC performance of medium quality speech from the GSM network. The original speech files were corrupted with babble noise at SNR = 15 dB and 10-15% frame loss. The Background set contained medium-quality coded speech (i.e., matched scenario).

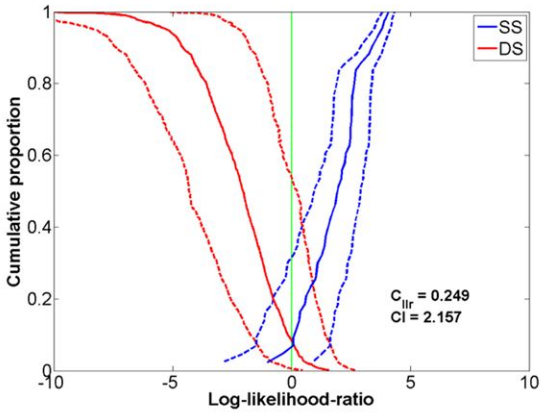

Figure 6: Tippett plot showing the FVC performance of medium-quality speech from the CDMA network. The original speech files were corrupted with babble noise at SNR = 15 dB and 10-15% frame loss. The Background set contained medium-quality coded speech (i.e., matched scenario).

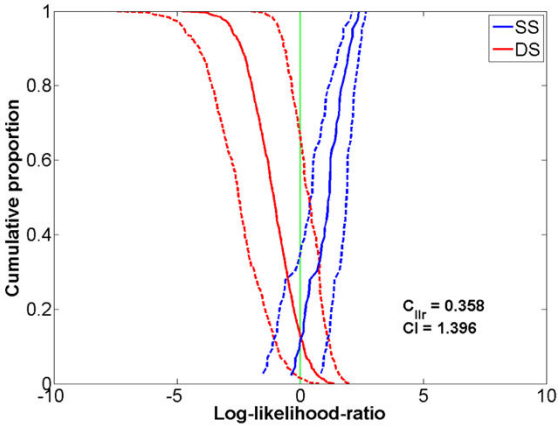

Figure 7: Tippett plot showing the FVC performance of medium quality speech from the CDMA network. The original speech files were corrupted with babble noise at SNR = 15 dB and 10-15% frame loss. The Background set contained un-coded speech (i.e., mismatched scenario).

The losses in C_{llr} were further examined using APE plots for no-coded speech and for both networks for the case of coded speech with babble noise at SNR = 15dB. Figure 8 shows the APE plot for matched conditions for coded speech and Figure 9 shows it for mismatched. Firstly, in all cases including no-coded speech, C_{llr} is dominated by discrimination loss (C_{llrmin}). This is marginally higher for the CDMA network compared to the GSM network.

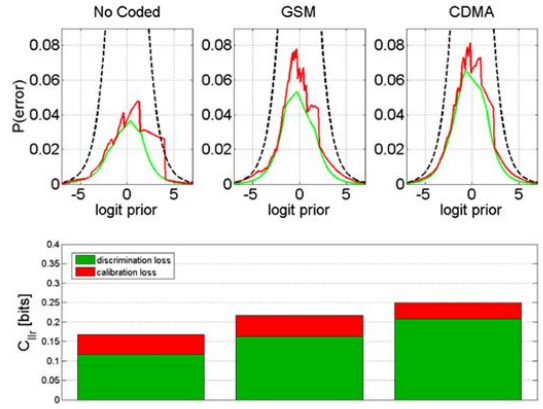

Figure 8: APE-plot for various speech conditions in GSM and CDMA networks at SNR = 15 dB and 10-15% frame loss. The Background set contained medium-quality coded speech (i.e., matched scenario).

This is an expected result likely due to the presence of NS in the CDMA network which can have the unfortunate consequence of removing part of the original speech, as previously discussed in earlier section. Both networks have slightly worse discrimination loss than no-coded speech under matched conditions. In the case of mismatched

conditions, the increase in discrimination loss was quite significant for both networks. Losses due to calibration were quite similar for most of the investigated scenarios except for the CDMA network under mismatched conditions. In this case, a small increase in calibration loss is observed, but overall the impacts on calibration loss are inconclusive.

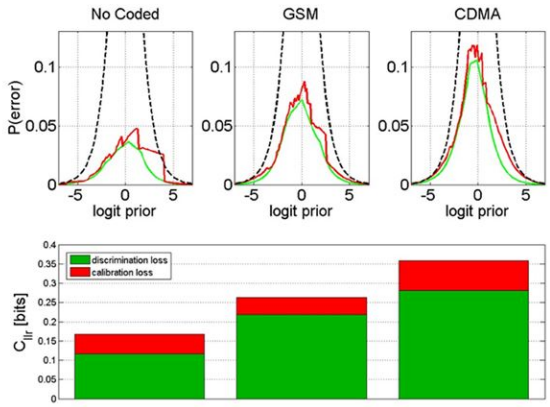

Figure 9: APE-plot for various speech conditions in GSM and CDMA networks at SNR = 15 dB and 10-15% frame loss. The Background set contained un-coded speech (i.e., mismatched scenario).

Conclusion

The reliability and accuracy of a FVC analysis using GSM- and CDMA-coded speech has been investigated in this paper. A number of key aspects have been incorporated to reflect realistic transmission scenarios in both networks, namely DRC, FL and BN. Average channel conditions have been assumed for both networks and performance has been investigated for different levels of BN.

FVC accuracy has been shown to be consistently worse for coded speech than un-coded irrespective of the network of origin. Comparing the two networks in this regard, CDMA-coded speech has been shown to give worse FVC accuracy than GSM-coded speech, particularly when BN levels are high. It is conjectured that this latter observation is linked to the NS process implemented in the CDMA network. For both networks, FVC accuracy worsens as BN levels increase, which is an expected result.

Our results have shown that better FVC accuracy results if an analysis is undertaken using matched conditions (i.e., the Background set is coded similarly to the Development and Testing sets), than unmatched (i.e., the Background set uses no-coded speech, whereas the Development and Testing sets use coded speech). This is particularly so for CDMA-coded speech and is an important finding for the forensic scientist because this is one aspect that is under their control.

With respect to FVC reliability, coding has been shown to actually improve this aspect, which is an unexpected result. We have conjectured that this improvement may be linked to the quantization processes associated with the coding algorithms which may have the effect of lessening parameter variations and thereby lessening intra-speaker variation. For both matched and mismatched conditions, CDMA-coded speech has been shown to result in generally better FVC reliability than GSM-coded. As far as BN is concerned, our results have shown no clear trend between SNR and FVC reliability. Finally, mismatched analysis generally results in better FVC reliability.

References

1. Morrison GS (2010) Forensic voice comparison. Expert Evidence 40: 1-105.
2. Alzqhoul EA, Nair BB, Guillemin BJ (2012) Speech Handling Mechanisms of Mobile Phone Networks and Their Potential Impact on Forensic Voice Analysis. SST, Sydney, Australia.
3. 3GPP TS 26.090 V11.0.0 (2016) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions.
4. Alzqhoul EA, Nair BB, Guillemin BJ (2015) Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison. Science & Justice 55: 363-374.
5. Alzqhoul EA, Nair BB, Guillemin BJ (2014) An Alternative Approach for Investigating the Impact of Mobile Phone Technology on Speech. World Congress on Engineering and Computer Science, San Fransisco, USA.
6. 3GPP2-S0018-D (2013) Minimum Performance Specification for the Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, and 73 for Wideband Spread Spectrum Digital Systems.
7. Nokia (2004) Guidelines for practical implementation of AMR in Nokia's Network element. General description.
8. 3GPP2-EVRC (2007) E.V.R. Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems General description.
9. 3GPP TS 45.009 (2008-2012) 3rd Generation Partnership Project; Technical specification Group GSM/EDGE Radio Access Network; Link adaptation.
10. 3GPP TS 26.091 (2012) 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.
11. Networks D (2012) Voice Quality Solutions for Wireless Networks.
12. 3GPP TS 06.11 (1999) Substitution and Muting of Lost Frames for Full Rate Speech Channels. http://www.3gpp.org/,
13. G. T. 26.077 (2012) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Minimum performance requirements for Noise Suppresser; Application to the Adaptive Multi-Rate (AMR) speech encoder.
14. G. T. 26. 978 (2012) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Results of the Adaptive Multi-Rate (AMR) noise suppression selection phase.
15. Punter S (2013) Southern Ontario Cell Phone Page.
16. Rose P (2003) The technical comparison of forensic voice samples, Expert Evidence.
17. Rose P (2002) Forensic speaker identification: CRC Press.
18. Rodriguez JG, Rose P, Ramos D, Toledano DT, Garcia JO (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. Audio, Speech, and Language Processing, IEEE Transactions 15: 2104-2115.
19. Aitken CG, Lucy D (2004) Evaluation of trace evidence in the form of multivariate data. Journal of the Royal Statistical Society. Series C (Applied Statistics) 53: 109-122.
20. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digital signal processing 10: 19-41.
21. Nair BB, Alzqhoul EA, Guillemin BJ (2012) A new approach to computing likelihood ratios based on principal component analysis. UNSW Forensic Speech Science Conference, Sydney, Australia.

22. Kim W, Hansen JH (2009) Feature compensation in the cepstral domain employing model combination. Speech Communication 51: 83-96.
23. Milner B, Darch J, Vaseghi S (2008) Applying noise compensation methods to robustly predict acoustic speech features from MFCC vectors in noise. IEEE International Conference pp: 3945-3948.
24. Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification.
25. Gold E, French P (2011) International practices in forensic speaker comparison. International Journal of Speech Language and the Law 18: 293-307.
26. Morrison GS (2009) Forensic voice comparison and the paradigm shift. Science & Justice 49: 298-308.
27. Morrison GS (2008) Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English/Ai. International Journal of Speech Language and the Law 15: 249-266.
28. Shlens J (2005) A tutorial on principal component analysis. Systems Neurobiology Laboratory, University of California: San Diego.
29. Jackson JE (2005) A user's guide to principal components.
30. Jolliffe I (2005) Principal component analysis.
31. Aitken CG, Taroni F (2004) Statistics and the evaluation of evidence for forensic scientists. 2: 540.
32. Morrison GS (2011) Measuring the validity and reliability of forensic likelihood-ratio systems. Science & Justice 51: 91-98.
33. Meuwly D, Drygajlo A (2001) Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). A Speaker Odyssey, The Speaker Recognition Workshop.
34. Brümmer N, Du Preez J (2006) Application-independent evaluation of speaker detection. Computer Speech & Language 20: 230-275.
35. Brummer N, Burget L, Cernocky JH, Glembek O, Grezl F, et al. (2006) Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation. Audio, Speech, and Language Processing, IEEE Transactions 15: 2072-2084.
36. Morrison GS, Thiruvaran T, Epps J (2010) Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. Proceedings of Odyssey pp: 63-70.
37. Morrison GS, Zhang C, Rose P (2011) An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic Science International 208: 59-65.
38. Messer K, Matas J, Kittler K, Luettin J, Maitre G (1999) XM2VTSDB: The extended M2VTS database. Second international conference on audio and video-based biometric person authentication pp: 965-966.
39. (2012) GoldWave.
40. (2011) Wavesurfer.
41. Guillemin BJ, Watson C (2008) Impact of the GSM Mobile Phone Network on the Speech Signal–Some Preliminary Findings. International Journal of Speech Language and the Law 15: 193-218.
42. Guillemin BJ, Watson C (2006) Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification. Proceedings of the 11th Australian International Conference on Speech Science & Technology.
43. Castro DR, Rodriguez G, Garcia JO (2006) Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. Speaker and Language Recognition Workshop, IEEE Odyssey pp: 1-8.
44. Alzqhoul EA, Nair BB, Guillemin BJ (2014) Comparison between Speech Parameters for Forensic Voice Comparison Using Mobile Phone Speech.
45. Rabiner LR, Schafer RW (2009) Theory and application of digital speech processing: Pearson.
46. Nair BB, Alzqhoul EA, Guillemin BJ (2014) Comparison between Mel-frequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA.
47. Kuhn V (1997) Applying list output Viterbi algorithms to a GSM-based mobile cellular radio system. IEEE 6th International Conference 2: 878-882.
48. Krishnamurthy N, Hansen JH (2009) Babble noise: modeling, analysis, and applications. Audio, Speech, and Language Processing, IEEE Transactions 17: 1394-1407.
49. ITU-T (2011) Objective measurement of active speech level ITU-T recommendation pp: 56.
50. (2013) Sound Effects.
51. Nair B, Alzqhoul E, and Guillemin BJ (2014) Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis. International Journal of Speech, Language and the Law 21: 83-112.