

Impact of Frame Loss Aspects of Mobile Phone Networks on Forensic Voice Comparison

Balamurali BT Nair^{1,2*}, Esam AS Alzqhouli^{1,2} and Bernard J Guillemain^{1,2}

¹Forensic and Biometrics Research Group (FaB), The University of Auckland, Auckland, New Zealand

²Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand

Abstract

The analysis of mobile phone speech recordings can play an important role in criminal trials. However it may be erroneously assumed that all mobile phone technologies, such as the Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA), are similar in their potential impact on the speech signal. In fact these technologies differ significantly in their design and internal operation. This study investigates the impact of an important aspect of these networks, namely Frame Loss (FL), on the results of a forensic voice comparison undertaken using a Bayesian likelihood ratio framework. For both networks, whenever a frame is lost or irrecoverably corrupted, it is synthetically replaced at the receiving end using a history of past good speech frames. Sophisticated mechanisms have been put in place to minimize any resulting artefacts in the recovered speech. In terms of accuracy, FL with GSM-coded speech is shown to worsen same-speaker comparisons, but improve different-speaker comparisons. In terms of precision, FL negatively impacts both sets of comparisons. With CDMA-coded speech, FL is shown to negatively impact the accuracy of both same- and different-speaker comparisons. However, surprisingly, FL is shown to improve the precision of both sets.

Keywords: GSM; CDMA; Forensic voice comparison; Likelihood ratio; Frame loss; Frame error rate

Abbreviations: AMR: Adaptive Multi Rate; APE: Applied Probability of Error; BN: Background Noise at the Transmitting End; CDMA: Code Division Multiple Access; CELP: Code Excited Linear Prediction; CI: Credible Interval; C_{llr} : Log-Likelihood-Ratio Cost DRC: Dynamic Rate Coding; EVRC: Enhanced Variable Rate Codec; FER: Frame Error Rate; FL: Frame Loss Mechanism; FVC: Forensic Voice Comparison; GMM-UBM: Gaussian Mixture Model-Universal Background Model; GSM: Global System for Mobile Communications; LR: Likelihood Ratio; LLR: Log-Likelihood Ratio; MFCCs: Mel-Frequency Cepstral Coefficients; MOS: Mean Opinion Score; MVKD: Multivariate Kernel Density; OP: Anchor Operating Point; PCA: Principal Component Analysis; PCAKLR: Principle Component Analysis Kernel Likelihood Ratio; PESQ: Perceptual Evaluation of Speech Quality; PPP: Pitch Period Prototype

Introduction

Mobile phone recordings are often used as evidence in courts of law. Analysis of such recordings using a range of forensic voice comparison (FVC) techniques can assist the court in establishing the guilt or innocence of a suspect. Forensic speech scientists when undertaking such analysis may erroneously assume that all mobile phone networks impact the speech signal in a similar manner. The most widely used mobile phone technologies in use today are Global System for Mobile Communications (GSM) and Code Division Multiple Access (CDMA). There are three key aspects of these networks which can directly impact the speech signal and thus the outcome of a FVC analysis: (i) dynamic rate coding (DRC), (ii) strategies for handling lost or corrupted frames (FL), and (iii) strategies for overcoming the effects of background noise at the transmitting end (BN). In [1] we examined the 1st of these. This paper directly follows on from that work and examines the 2nd factor, the impact of FL in these two networks.

In mobile phone networks speech is coded into 20 ms frames. The wireless channel associated with these networks can often be quite poor, necessitating the need for innovative techniques to try and ensure reliable transmission. Notwithstanding this, the following

could happen to a transmitted frame: (i) it is lost, (ii) it is received, but in a corrupted state, or (iii) it is received without error. In the case of a corrupted frame, techniques such as convolutional coding [2,3] are used to try and correct for errors. If correction is not possible, the FL mechanism is initiated. For both networks this broadly involves replacing lost speech data with speech data from the past.

Much of the experimental methodology of this study is the same as that of our previous DRC study and the reader is referred to that paper for an in depth explanation and justification of our approach [1]. We again use the Bayesian likelihood ratio (LR) framework for the evaluation of speech forensic evidence. A number of methods have been proposed for evaluating speech evidence in the FVC arena, such as Gaussian mixture model universal background model (GMM-UBM) [4,5], multivariate kernel density (MVKD) [4,6] and principal component analysis kernel likelihood ratio (PCAKLR) [7]. Each of these computes a LR, which is a ratio of probabilities. The numerator of the LR is the probability of the evidence given the prosecution hypothesis; the denominator is the probability of the evidence given the defence hypothesis. GMM-UBM has been primarily designed for data-stream-based analysis scenarios, whereas MVKD and PCAKLR are primarily designed for token-based analysis scenarios [8]. The difference between MVKD and PCAKLR is principally in respect to the number of parameters that can be handled, this being 3-4 in the case of MVKD [6], and much larger than this in the case of PCAKLR [9]. Given

***Corresponding author:** Balamurali BT Nair, Forensic and Biometrics Research Group (FaB), The University of Auckland, Auckland, New Zealand, Tel: +6493737599; E-mail: bbah005@aucklanduni.ac.nz

Received October 29, 2015; **Accepted** November 26, 2015; **Published** November 30, 2015

Citation: Nair BBT, Alzqhouli EAS, Guillemain BJ (2015) Impact of Frame Loss Aspects of Mobile Phone Networks on Forensic Voice Comparison. *Sensor Netw Data Commun* 4: 131. doi:10.4172/2090-4886.1000131

Copyright: © 2015 Nair BBT, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

that, as for our previous study, we use vowel tokens for the experiments reported here, these being represented by 23 Mel-Frequency Cepstral Coefficients (MFCCs), PCAKLR has been chosen for computing LRs.

To quantify the performance of a FVC experiment, we use here the same tools used in our DRC study, namely log-likelihood-ratio cost (C_{lr}), Tippett plots, applied probability of error (APE) plots, and credible interval (CI). The reader is again referred to our earlier paper for more details on these [1].

The remainder of this paper is structured as follows. The FL mechanisms for both the GSM and CDMA networks are discussed first in great detail, followed by our experimental methodology to study the impact of these on FVC. We then present our results and conclusions.

FL Mechanisms in the GSM and CDMA Networks

With both the GSM and CDMA networks, lost or irrecoverably corrupted frames are replaced with synthetically generated frames using speech data derived from the past, a process which is implemented by the decoding section of the speech codec used in the network. The most widely used speech codecs in the GSM and CDMA networks are the adaptive multi rate (AMR) codec and enhanced variable rate codec (EVRC), respectively. For both networks, if successive frames are lost, the codec will continue replacing those, while at the same time gradually decreasing the output level until silence results, a process called muting [10]. A maximum of 16 successive frames (i.e., 320 ms) could be replaced in this manner before silence results [11,12].

From the perspective of a FVC, the automatic replacement of lost frames with synthetically generated frames is clearly of concern, unless their occurrence can be detected *a priori* and the synthetically generated sections excluded from an analysis. But the codecs have been designed with speech quality in mind and sophisticated strategies have been incorporated, such as smoothing out any abrupt amplitude transitions from one speech frame to another, to minimize or even eliminate any resulting perceptual artefacts. So effective are these strategies that subsequent detection of the FL process from the received speech signal is likely to be very difficult, if not impossible.

We believe it is important for forensic speech scientists to clearly appreciate that with mobile phone speech much of the decoded speech waveform could be artificially generated, and that this must necessarily impact upon the confidence they ascribe to any of their analysis findings. With the intention of convincingly making this point, the following discussion is deliberately detailed. However in reality it is not the specifics of the process that the forensic scientist needs to understand, but rather they need to have an overall appreciation of how much of the speech waveform, and in what respects, it might have been changed during transmission.

GSM FL mechanism

The AMR codec processes speech frames using code excited linear prediction (CELP) into 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 [13,14]. This multi-bit-rate capability is designed to allow the GSM network to use available transmission bandwidth as efficiently as possible in response to changing channel conditions [15].

The AMR FL mechanism is quite sophisticated [16]. The example of Figure 1 is intended to illustrate some of the key features of this process. Figure 1a shows a sequence of seven received speech frames. Four of these have been received without error and are therefore 'Good' (labelled with a superscript G), while the remaining three,

having been identified as containing irrecoverable errors, are therefore 'Bad' (labelled with a superscript B in the figure). Figure 1b shows the resulting speech frames that would be used to generate the decoded speech waveform. With this example, in order to convey the broad aspects of a process which in reality is quite complicated, we draw a distinction between data in a frame that could be classified as speech data (i.e., spectral shaping, voiced/voiceless classification, pitch, etc.) and data related to amplitude. We first consider how speech data gets impacted, then amplitude data.

The first two received frames, Frames 1 and 2, being 'Good', remain unchanged. The speech data of Frame 3, being 'Bad', is thrown away and replaced by speech data derived from the last 'Good' frame, namely Frame 2. The result is an artificially generated frame to replace the 'Bad' Frame 3. There is also an amplitude adjustment process associated with the generation of such frames, namely a gain reduction, as will be described below. In Figure 1b this new Frame 3 is labelled $3^R(2/)$, where the superscript R indicates a replaced frame, 2 indicates that its speech data has been derived from Frame 2, and '/' indicates that an amplitude adjustment has been applied. A similar process happens for the 'Bad' Frame 4, its speech data being derived from the synthetically-generated Frame 3, but with a further level of amplitude reduction. Thus in Figure 1b the new Frame 4 is labelled $4^R(3//)$ to indicate that its speech data has been derived from Frame 3, but now with two levels of amplitude adjustment. Frame 5 is 'Good', so its speech data is retained, but because it was preceded by a 'Bad' frame, its amplitude is also adjusted in an attempt to minimise amplitude discontinuities. Thus in Figure 1b it is labelled $5^R(5/)$. Frame 6 is 'Good', and given that it was preceded by a 'Good' frame, it is used without modification. Frame 7 is 'Bad', so its speech data is derived from Frame 6, but with an amplitude adjustment. It is therefore labelled as $7^R(6/)$.

In rather simplistic terms, the amplitude adjustment process associated with the AMR's FL mechanism works as follows. Each 20 ms frame is segmented into four 5 ms sub-frames, each with its own amplitude. Since Frames 1 and 2 are 'Good', the amplitudes of their sub-frames remain unchanged. In order to determine the amplitude, β_{31} , of the 1st sub-frame of the new Frame 3 (i.e., Frame labeled $3(2/)$), the median value β_{median} of the amplitudes of the previous five sub-frames is determined. For illustration purposes, these amplitudes will be referred to as β_{14} , β_{21} , β_{22} , β_{23} , and β_{24} , where β_{xy} is the amplitude of the yth sub-frame of Frame x. If $\beta_{24} \leq \beta_{median}$ then $\beta_{31} = \beta_{24} \times \alpha$, otherwise $\beta_{31} = \beta_{median} \times \alpha$, where α is some attenuation factor. (Note that the value of α is not constant, but changes dependent upon such factors as the sequence of 'Good' and 'Bad' frames received) [16]. An identical process

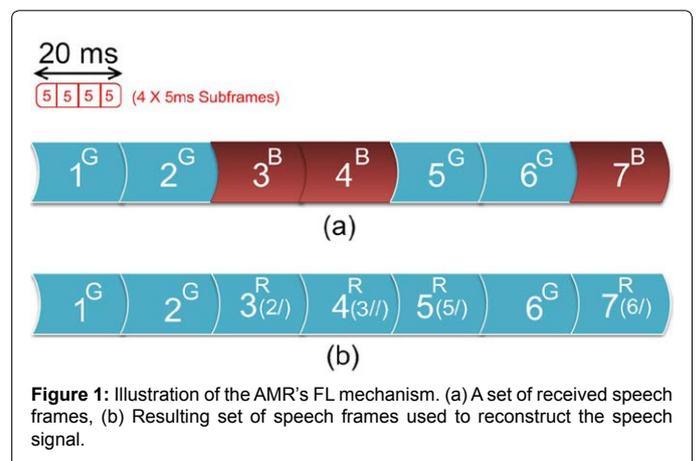


Figure 1: Illustration of the AMR's FL mechanism. (a) A set of received speech frames, (b) Resulting set of speech frames used to reconstruct the speech signal.

is used to determine the amplitudes of the remaining sub-frames of the new Frame 3, as well as of all of the sub-frames in the new Frame 4. So for instance, in respect to determining the amplitude, β_{32} , of the 2nd sub-frame of Frame 3, the β_{median} value used is determined from β_{21} , β_{22} , β_{23} , β_{24} and β_{31} , where β_{31} has the same meaning as before, namely the amplitude of the 1st sub-frame of the new Frame 3 just determined.

The process in respect to deciding on the amplitudes of the sub-frames of the new Frame 5, namely β_{51} , β_{52} , β_{53} and β_{54} , is different to the one just described for the sub-frames of the new Frames 3 and 4 because Frame 5 was a ‘Good’ frame preceded by a ‘Bad’ frame. So the only changes made to it are in respect to the amplitudes of its sub-frames, and this for the sole reason of minimising discontinuities.

The process here is based on a single comparison between the amplitude of each of the sub-frames of Frame 5 and the amplitude of the previous ‘Good’ sub-frame received (relative to it). So in respect to deciding on the amplitude of the 1st sub-frame of Frame 5, the previous ‘Good’ sub-frame received (again relative to it) is the 4th sub-frame of Frame 2. If $\beta_{51} \leq \beta_{24}$, then β_{51} remains unchanged, else $\beta_{51} = \beta_{24}$. Similarly a value for β_{52} is decided by comparing it to β_{51} (i.e., the value just decided in the previous step). Specifically, if $\beta_{52} \leq \beta_{51}$, then β_{52} remains unchanged, else $\beta_{52} = \beta_{51}$. This recursive process is repeated for the remaining two sub-frames of Frame 5. Frame 6 remains unchanged in the FL process, and this is in all respects including the amplitudes of its sub-frames. The amplitudes of the sub-frames of the final replaced frame, namely Frame 7, are determined in exactly the same manner as for Frames 3 and 4. So, for example β_{median} , the value used for determining β_{71} will be based on β_{34} , β_{61} , β_{62} , β_{63} and β_{64} [16].

It can be seen from this example that though there were four ‘Good’ frames and three ‘Bad’ frames in the received seven-frame sequence of Figure 1a, only three of these ‘Good’ frames, along with four artificially generated replacement frames, have been used to produce the decoded speech waveform. But the bit rates for all frames, whether ‘Good’ and therefore unchanged, or artificially generated to replace ‘Bad’ frames, remain the same. It is also evident from this example that a considerable degree of sophistication has been designed into the AMR’s FL process to mask any resulting perceptual artefacts, making its subsequent detection from the decoded speech waveform, in all likelihood, impossible.

CDMA FL mechanism

Before discussing the specifics of the FL strategy implemented by the EVRC decoder, it is necessary to give a brief overview of the codec itself. It operates in one of three modes, referred to as anchor operating points (OP), namely OP0, OP1, and OP2. The selection of a particular mode is made by the network according to the number of users accessing it. Once selected, the mode then defines the general behaviour of the codec as well as playing a role in determining the source coding bit rate for each speech frame. Upon selecting an OP, a speech frame is categorised as either voiced, voiceless, transient or silence. A source coding bit rate is then selected accordingly [17]. The codec produces output frames at one of four source coding bit rates, namely 8.55, 4, 2 and 0.8 kbps, with the latter being used to code silence frames. It also uses a number of coding techniques such as code excited linear prediction (CELP), pitch prototype period (PPP) and silence coder, these being selected for an individual frame on the basis of its speech category and the OP chosen.

The EVRC’s FL mechanism also involves replacing ‘Bad’ frames with ‘Good’ frames using speech data from the past, but unlike with the AMR codec, an artificially created ‘Good’ frame is not necessarily at the

same bit rate as the ‘Bad’ frame it replaces. Usually it is set to the highest bit rate of 8.55 kbps. To illustrate the various aspects of the EVRC’s FL strategy [17], we again use a similar example as for the AMR codec. We also again draw a distinction between speech data and amplitude data in a frame. But, unlike for the AMR codec, the EVRC’s FL process is much simpler, so we discuss the handling of speech data and amplitude data at the same time.

Figure 2a shows a sequence of received frames, four ‘Good’ and three ‘Bad’. The superscripts G and B associated with individual frames have the same meaning as before. Subscripts refer to the speech frame type such as silence (identified with S) or active speech (identified with an associated bit rate, namely 2, 4 or 8.55 kbps). Figure 2b shows the resulting speech frames that would be used to generate the decoded speech waveform. Again, the superscript R associated with an individual frame identifies it as a replacement frame and a subscript has the same meaning as in Figure 2a. Frames 1 and 2 are both ‘Good’ and therefore remain unchanged. Frame 3 is ‘Bad’ and is replaced by a synthetically-generated frame at a bit rate of 8.55 kbps. Essentially the speech data used in the new Frame 3 is the same as that in the last ‘Good’ speech frame, namely Frame 2, except for a possible modification needed to correct for any change in bit rate between the two frames. If the bit rate of Frame 2 was 8.55 kbps, then this will be used for the new Frame 3. If the bit rate for Frame 2 was either 2 or 4 kbps, a sophisticated bandwidth expansion of its speech data is performed to match the higher bit rate of the new Frame 3. As far as amplitude data for the new Frame 3 is concerned, this is made the same as for Frame 2. Thus the new Frame 3 in Figure 2b is identified as $3_{8.55}^R(2)$.

Frame 4 is also ‘Bad’ and its speech data would be replaced in an identical manner to Frame 3 (i.e., based on the speech data from the last ‘Good’ frame, namely Frame 2, but again with a possible bandwidth expansion). However, unlike for the new Frame 3, there would be an associated reduction in amplitude by a factor of 0.75 because Frame 4 is the second ‘Bad’ frame in a sequence. Thus the new Frame 4 in Figure 2b is identified as $4_{8.55}^R(2/)$. (Note: if a sequence of frames is ‘Bad’, the same process would be repeated, but with the amplitude of all subsequent replaced frames being reduced by a factor of $(0.75)^{N-1}$, where N is the consecutive ‘Bad’ frame number ($N \geq 2$).) Frame 5 is ‘Good’, so remains essentially unchanged, except for its associated pitch parameter. Again with the goal of minimising discontinuities in the recovered speech signal, in this case in respect to pitch, the pitch information of Frame 5 would be altered to become essentially

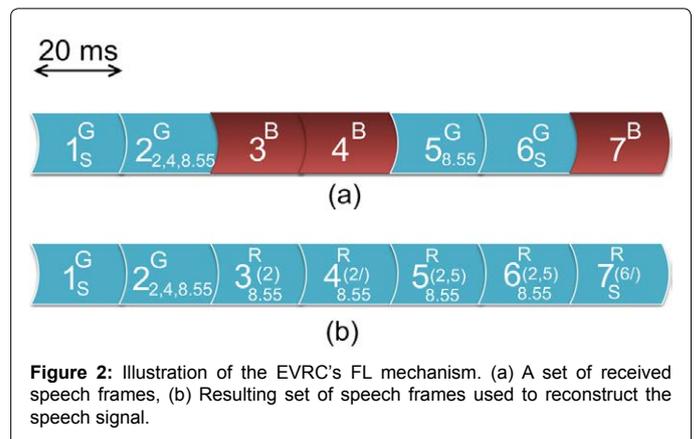


Figure 2: Illustration of the EVRC’s FL mechanism. (a) A set of received speech frames, (b) Resulting set of speech frames used to reconstruct the speech signal.

an interpolation between the pitch of Frame 2 (and thus of the new Frames 3 and 4 which would have the same pitch as Frame 2) and that of Frame 5. Thus in Figure 2b the new Frame 5 is labelled as $5_{8.55}^R$ (2,5) to indicate that it has derived its speech data from Frames 2 and 5. Frame 6, which is a silence frame, is also ‘Good’. However, one of the rules associated with the EVRC’s FL mechanism is that a silence frame cannot be preceded by a replaced frame that is high quality (i.e., a frame with a bit rate of 8.55 kbps). So Frame 6 is discarded and is replaced by a copy of the previous frame, namely the new Frame 5. It is thus labelled as $6_{8.55}^R$ (2,5) in Figure 2b. Finally, Frame 7 is ‘Bad’ and so is replaced by essentially a copy of the ‘Good’ Frame 6 that was received, the only modification being in respect to its amplitude, this being recalculated slightly differently to other frames using procedures outlined in [17] because it was preceded by a silence frame. The new Frame 7 then becomes 7_S^R (6/) in Figure 2b.

It can be seen from this example that though there were four ‘Good’ frames and three ‘Bad’ frames in the received seven-frame sequence of Figure 2a, this has resulted in only two of these ‘Good’ frames, together with five synthetically-generated frames, being used to generate the decoded speech waveform. It is also clear from this example that, as with the AMR codec, a considerable degree of sophistication has been incorporated into the EVRC’s FL mechanism, the underlying goal again being to conceal as far as possible, from a perceptual standpoint, that data has been lost or corrupted during transmission. The unfortunate consequence from the standpoint of a FVC analysis is that determining from the recovered speech signal when this process has occurred is likely to be very challenging, if not impossible.

Experimental Methodology

Speech database and speech parameters used

We used the same 130 male speakers from the XM2VTS database [18] as used in our DRC study [1], these being judged perceptually to have the same Southern British accent. The speakers were recorded on four different occasions separated by a one month interval. During each session each speaker read the following random digit sequences: 1. “zero one two three four five six seven eight nine”, and 2. “five zero six nine two eight one three seven four”. The speech files in the XM2VTS are sampled at 32 kHz with 16 bit digitization. We down-sampled these to 8 kHz to align with the input speech requirements of mobile codecs. Three recording sessions out of the four available have been used in our experiments. We focused on the three words, “nine”, “eight” and “three” from these recordings and extracted their corresponding vowel segments /aI/, /eI/ and /i/ (i.e., two diphthongs and a monophthong) using a combination of auditory and acoustic procedures [19]. In summary, three non-contemporaneous sessions have been used with three vowels per session and four tokens per vowel.

As for our previous DRC study, 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. (Note: the purpose of the Development set is to train the logistic regression-fusion system [20], the resulting weights of which are then used to combine LRs calculated from individual vowels for each comparison in the Testing set.) 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. Similarly, 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 (refer to Table 2 of our DRC paper [1]). The

Background set remained the same for all comparisons and contained two recording sessions for all 44 speakers. 23 MFCCs were then computed for coded speech under various conditions of FL using the same MFCC extraction process as used in our DRC study. C_{lr} values were calculated using the mean LRs for same- and different-speaker comparisons (note: two LRs were calculated for each same-speaker experiment and three LRs for each different-speaker experiment). CI was calculated by finding the variation in LR values (again, using two LRs for same-speaker comparisons and three LRs for different-speaker comparisons).

Strategies to understand the impact of FL

Our goal with this study was to, as far as possible, study the impact of FL on FVC in isolation to the other two factors in a mobile phone network that can impact speech quality, namely DRC and BN. Clearly any approach involving the transmission of speech across an actual network would not make this possible. So in this study we have again chosen to pass speech through software implementations of the codecs under investigation and introduce FL in a controlled manner, while endeavouring to disable both DRC and BN. Disabling DRC and BN is straight forward for the AMR codec. For the EVRC codec, however, though disabling BN is also straightforward, this is not so for DRC because the bit rate for a frame depends partly upon its classification (i.e., voiced, voiceless or transition) [17] and partly upon the codec mode (i.e., OP0, OP1 or OP2). Obviously a frame’s classification can’t be changed, but the mode can be constrained to one of the three.

Figure 3 shows a block diagram of the processing stages used in our experiments. The speech files were processed by each codec under two scenarios: one assuming no lost or corrupted frames and the other with speech frames lost or irrecoverably corrupted between the coder and decoder stages of the codec in some controlled manner, as will be discussed in the following section.

Simulating FL

The first aspect that needs to be considered when designing experiments of this kind is what level of frame loss, often referred to as frame error rate (FER), is typical of a real network. In mobile networks this parameter is constantly monitored during a call. When the FER exceeds in the region of 10 to 15% it is known that the overall voice quality degrades to a level where the mean opinion score (MOS) is less than about 2.9 [21]. Mobile network operators realise that such voice quality is unpleasant to the listener and they have therefore put procedures in place to automatically drop a call if this limit is reached.

In reality this monitoring of FER would be done over hundreds of frames corresponding to many seconds of speech. In our experiments, however, we have used vowel segments that are typically 12 to 15 frames

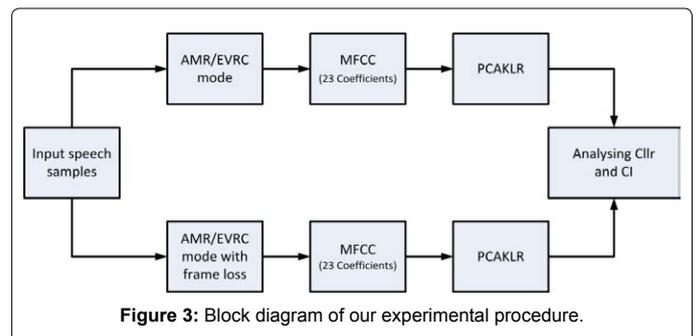


Figure 3: Block diagram of our experimental procedure.

in duration. In comparison to the duration of a vowel segment, the FER monitoring process described could be classified as a long-term statistical measure, and there would likely be short periods of time in which the actual FER was much higher. The question then arises as to whether this same upper value for FER of 10 to 15% is also appropriate for much shorter segments typical of vowels. To answer this question we conducted experiments where we examined the speech quality of vowels using PESQ [22] for a range of values of FER. In the interests of space we do not reproduce these experimental results here, but they showed that for vowel segments an FER in the region of 10 to 15% again translates into MOS values of the order of 2.9. So we used this same upper range for FER in our experiments as well. Given that the durations of our vowel segments were of the order of 12 to 15 frames, this FER rate translates into a maximum number of lost frames per vowel segment being typically one, or at most two. In the interests of investigating worst-case conditions, we have fixed the number of lost frames per vowel segment to two. For each vowel token, the locations of these lost frames has then been determined randomly according to a uniform distribution.

As shown in Figure 4, the speech files were coded at two different modes for each mobile network, these modes roughly translating into low and high quality speech coding. In the case of the GSM network, this was the 4.75 and 12.2 kbps modes, respectively, whereas in the CDMA network it was OP2 and OP0, respectively. For each mode, speech was coded twice, first without FL, then with it. The rationale behind conducting FVC experiments at two different speech qualities was to try and separate out the impact of speech coding quality from the impact of FL. It is important to note that the Background set used in these experiments contained coded speech at the specific mode being investigated, but without FL. This was done in an endeavour to minimise mismatch.

Results

Impact of FL on the decoded speech waveform

Before investigating the impact of FL on FVC performance, it is informative to examine how the temporal location of a lost frame, together with of course the associated FL corrective mechanism that it would have triggered, might impact upon the decoded speech waveform, both in terms of its temporal and spectral characteristics. To illustrate this, a set of time waveforms and spectrograms have been produced for a token of the vowel /aI/ coded with either the AMR or EVRC codec. A single lost frame has been introduced between the coded and decoded speech paths, but at three different temporal locations, namely at Frames 3, 4 and 5.

Figures 5 and 6 show the results for the AMR codec, with speech coded at 12.2 kbps. Figure 5a shows 180 ms of the time waveform of the vowel segment (i.e., 9 frames) without FL. Figures 5b-d show the resulting decoded speech waveform for Frames 3, 4 and 5 being lost, respectively. For the purpose of comparison, the amplitude of each of these time waveforms has been normalized to the maximum absolute value of the waveform in Figure 5a. Figure 6 shows the spectrograms corresponding to the time waveforms shown in Figure 5. Examination of both Figures 5 and 6 shows that the loss of a single frame can have quite an impact on all subsequent frames and that this impact depends very much on exactly which frame is lost.

The corresponding results for the EVRC are shown in Figures 7 and 8, with speech coded at OP0.

It is interesting to note that though it is exactly the same vowel segment that has been coded by both codecs, there are even differences between the resulting coded speech waveforms for the situation of no FL. As for the AMR codec, with the EVRC the loss of a single frame can have quite an impact on the subsequent decoded speech frames.

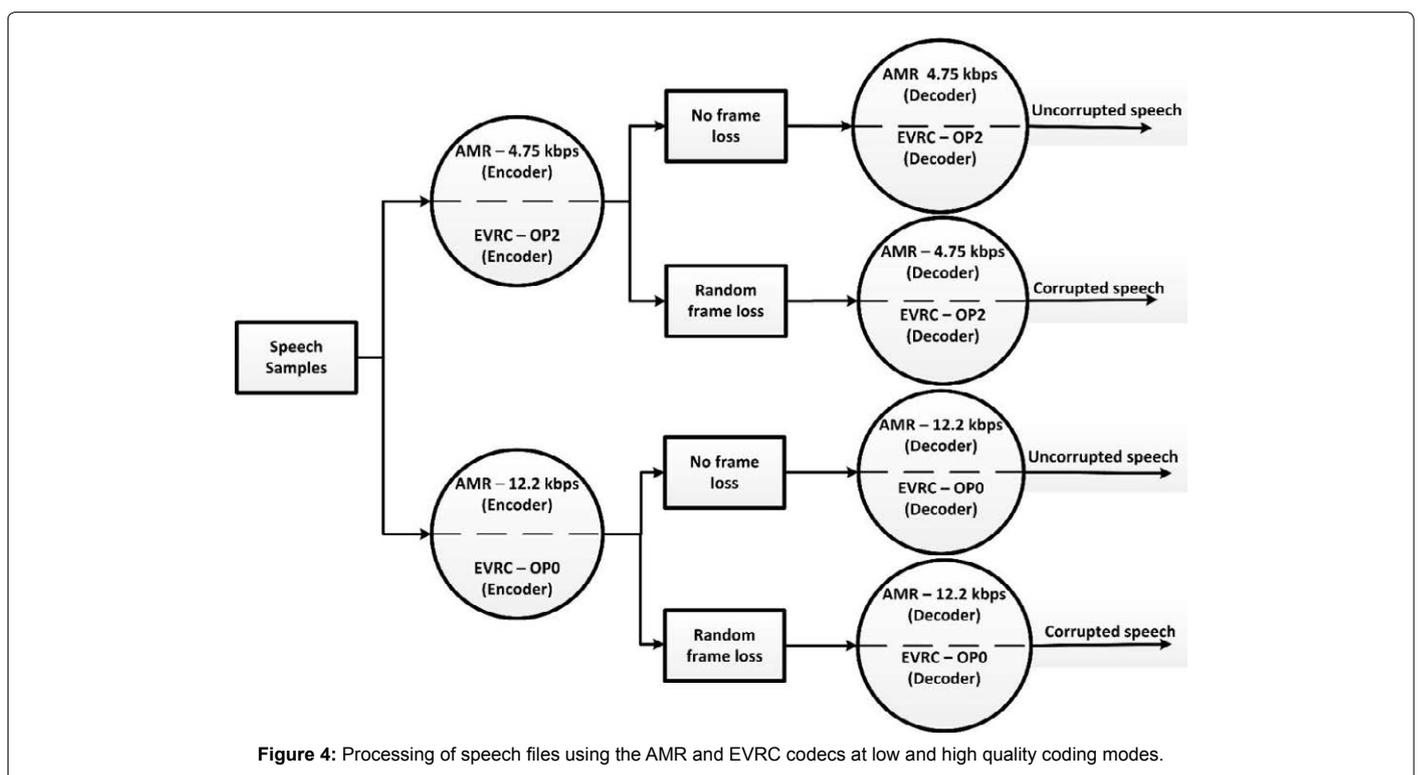
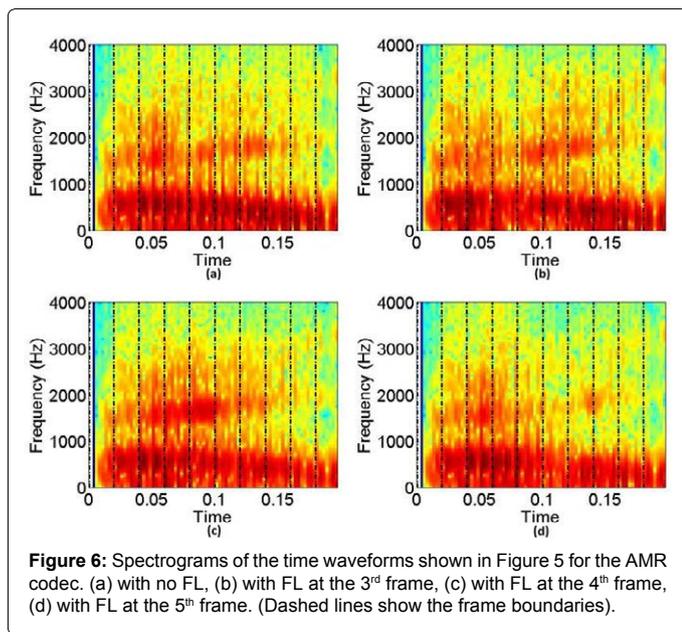
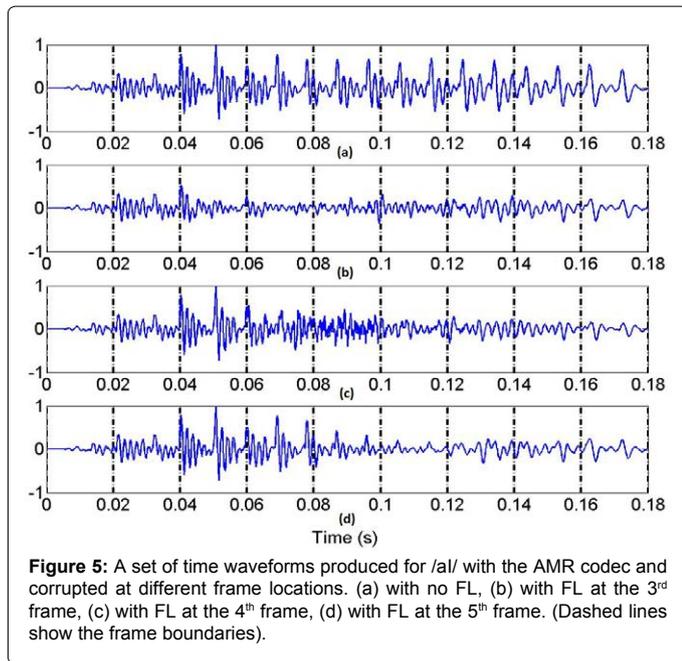


Figure 4: Processing of speech files using the AMR and EVRC codecs at low and high quality coding modes.



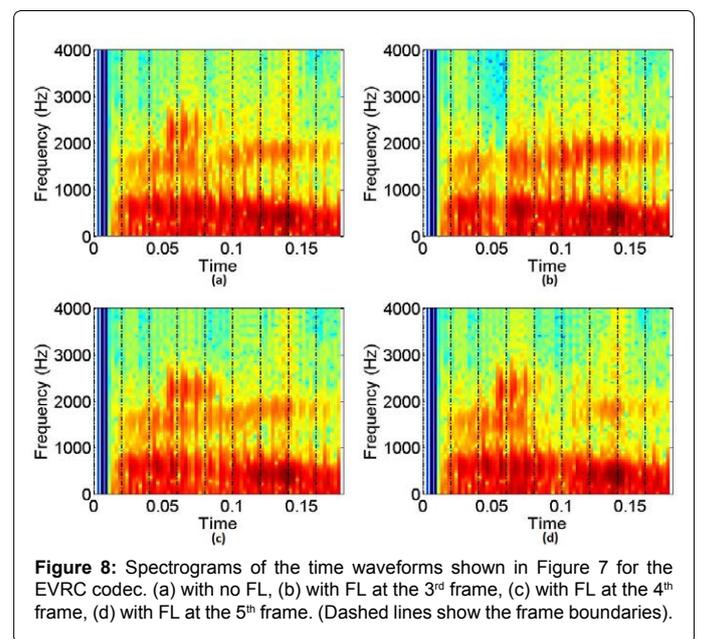
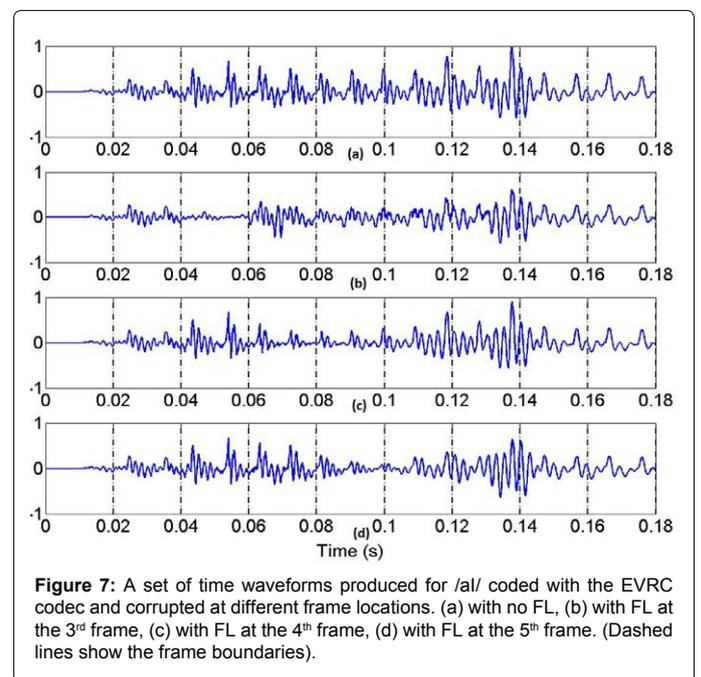
Impact of FL on FVC performance

This section presents results showing the impact on FVC performance arising from FL for both the AMR and EVRC codecs. Exactly two lost frames have been introduced into each vowel segment, their temporal locations being randomly determined according to a uniform distribution. LR values have been computed separately for each of the vowels /aI/, /eI/ and /i/ and the results then fused using logistic-regression fusion. The resulting FVC performance is shown in terms of C_{llr} , CI, Tippett plots and APE plots.

AMR codec: Figure 9 examines the impact of FL on FVC performance in terms of CI and C_{llr} for the AMR codec at 4.75 kbps and 12.2 kbps. Results are presented without and with FL for both cases. It is clear from these results that FL does have a negative impact upon

FVC performance, both in terms of accuracy (C_{llr}) and reliability (CI). Further, this impact is more severe for coded speech at the lower bit rate.

In order to investigate this latter aspect further, Figures 10 and 11 show Tippett plots for AMR-coded speech at 4.75 kbps, without and with FL, respectively. The corresponding results at 12.2 kbps are shown in Figures 12 and 13. The blue solid curve in these plots represents same-speaker comparison results and the red solid curve different-speaker comparison results. The dashed line on the either side of the blue and red curves represents the variation found in a particular LLR (i.e., $LLR \pm CI$). Considering first the results at 4.75 kbps, it is clear that a major impact of FL is on same-speaker classifications. The strength of both same-speaker and different speaker comparisons has increased



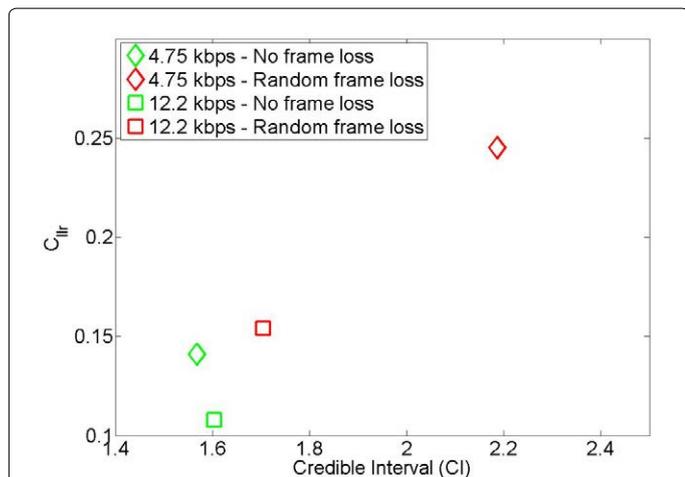


Figure 9: v.s. CI for AMR-coded speech at 4.75 kbps and 12.2 kbps without and with FL.

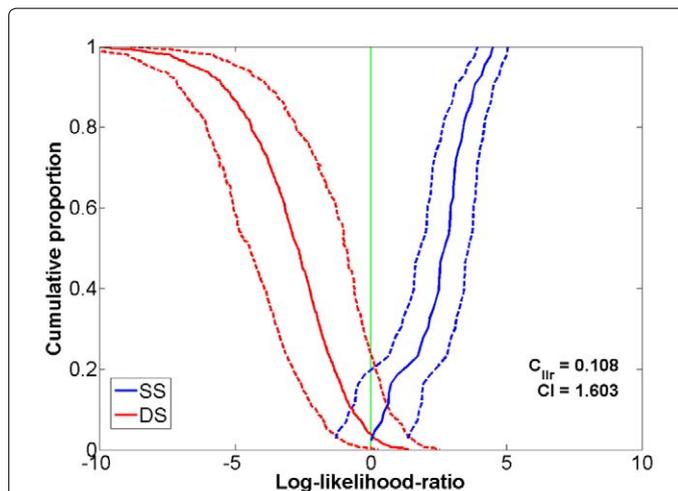


Figure 12: Tippett plot showing the performance of the AMR codec at 12.2 kbps without FL.

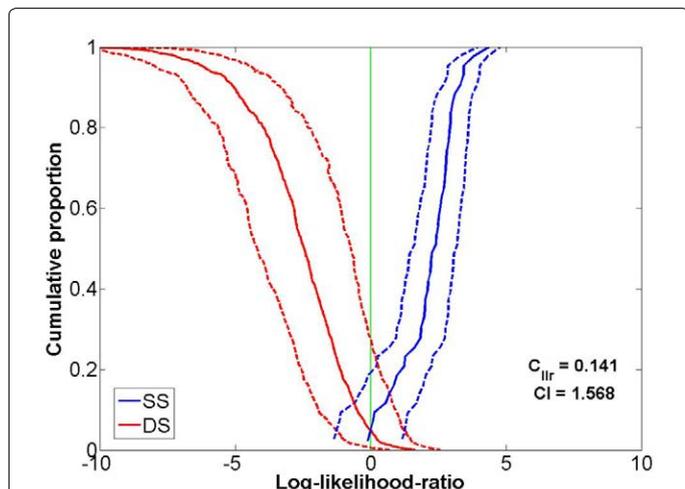


Figure 10: Tippett plot showing the performance of the AMR codec at 4.75 kbps without FL.

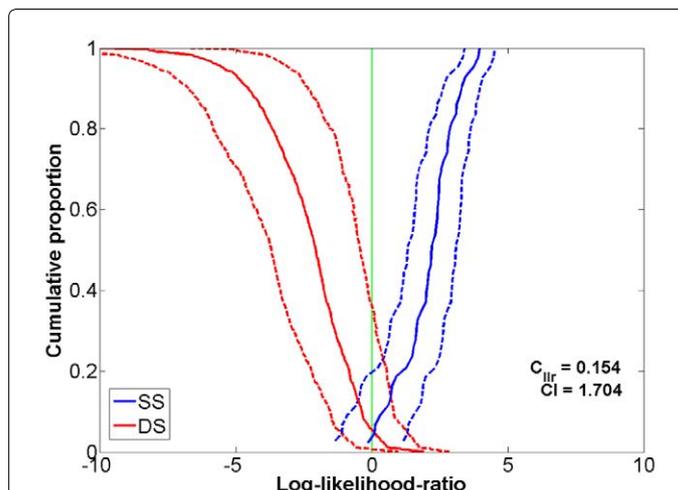


Figure 13: Tippett plot showing the performance of the AMR codec at 12.2 kbps with FL.

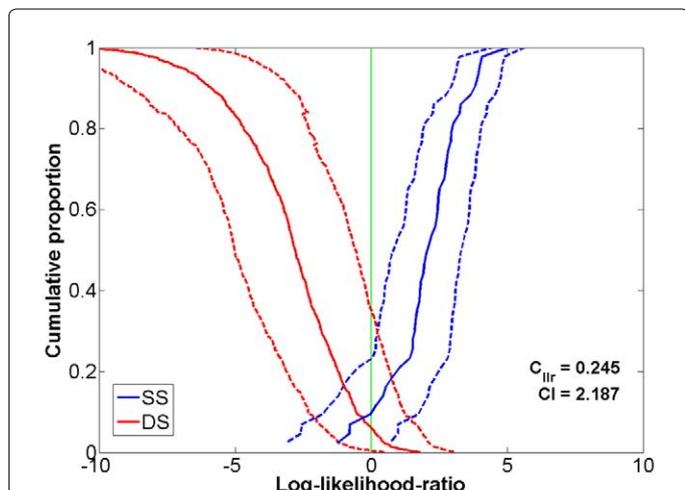
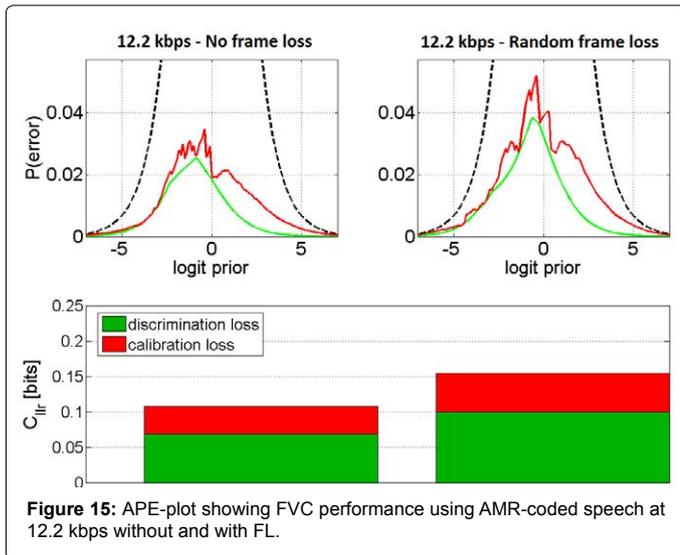
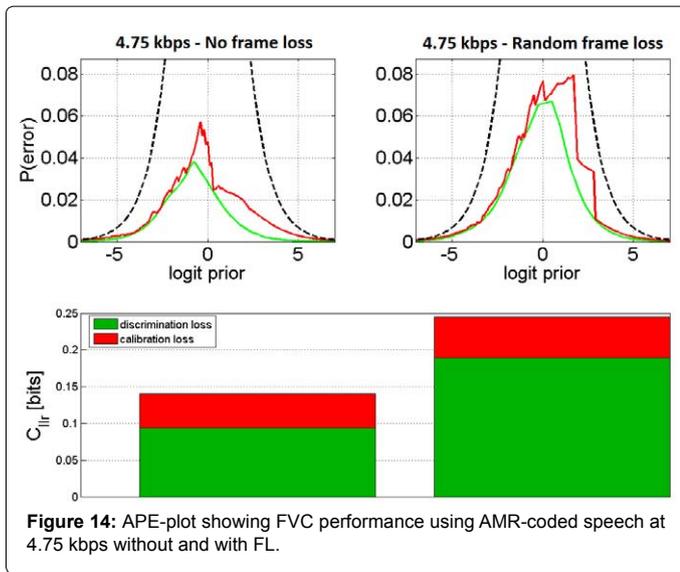


Figure 11: Tippett plot showing the performance of the AMR codec at 4.75 kbps with FL.

slightly, but importantly the number of same-speaker misclassifications has increased. Both of these findings are intuitively to be expected. In respect to different-speaker comparisons, the strength of these has improved slightly, which is again a finding one might expect. As far as reliability is concerned (i.e., CI), FL would appear to have a similar negative impact upon both same-speaker and different speaker comparisons. The results at 12.2 kbps (Figures 12 and 13) confirm that the impact of FL at the higher bit rate is fairly minimal, both in terms of same- and different-speaker comparisons.

To further understand what has contributed to the worsening of C_{lr} values as a result of FL, Figures 14 and 15 show APE-plots for the two cases of 4.75 kbps and 12.2 kbps, respectively. Considering first Figure 14 for 4.75 kbps, it is clear that FL has resulted in a significant increase in discrimination loss of almost 95%. Calibration loss has also increased, but only by about 20%. In the case of high bit rate coding (Figure 15), the calibration and discrimination loss components have both increased by about 40%.

EVRC Codec: Figure 16 examines the impact of FL on FVC performance in terms of CI and C_{lr} for the EVRC at OP2 (low quality



coding) and OP0 (high quality coding). For purposes of comparison, results are presented without and with FL for both cases. In respect to C_{lr} , the results for the EVRC are very similar to those for the AMR codec, namely, FL negatively impacts FVC accuracy and this is worse for low quality speech coding. Unlike for the AMR codec, however, for the EVRC for both low and high quality speech coding the CI has improved as a result of FL. Why this might be so is not clear at this stage.

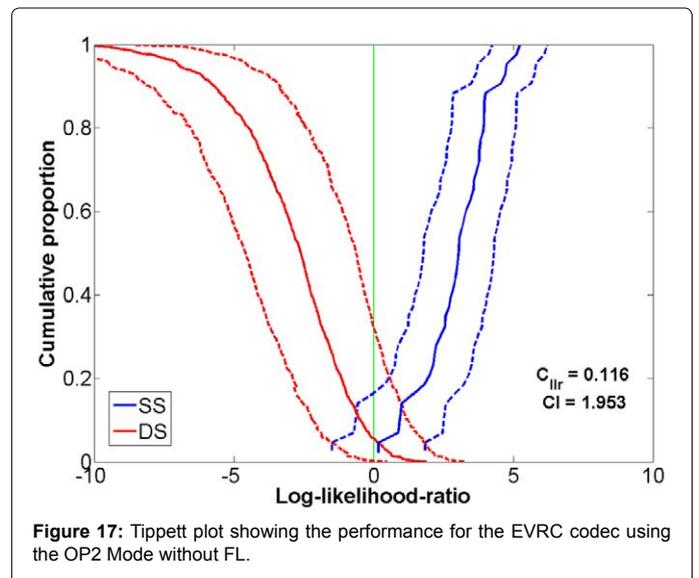
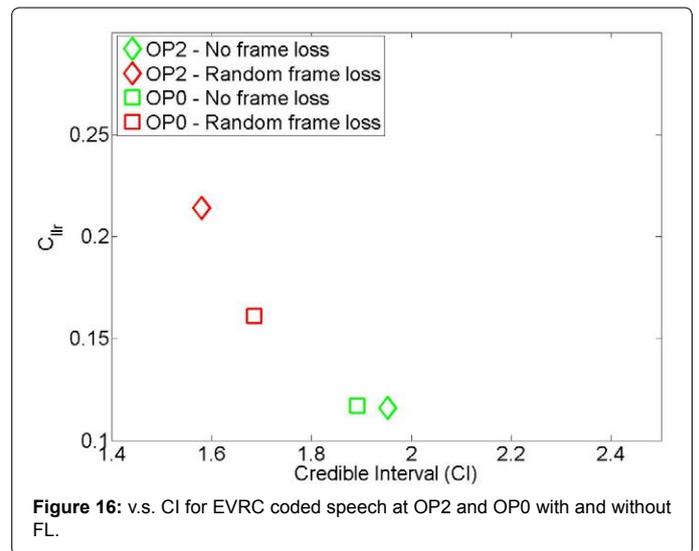
To further understand the degradation in C_{lr} values, Tippett plots are shown for OP2 without and with FL (Figures 17 and 18, respectively) and OP0 without and with FL (Figures 19 and 20, respectively). The first observation from these figures is that FL has negatively impacted both same- and different-speaker classifications, but this is less at the higher quality coding. Secondly, it has caused both same- and different-speaker misclassifications to increase, though for high quality coding this increase is minimal. As far as CI is concerned, Figures 17-20 confirm the previous finding for the AMR codec, namely FL has caused this aspect to improve.

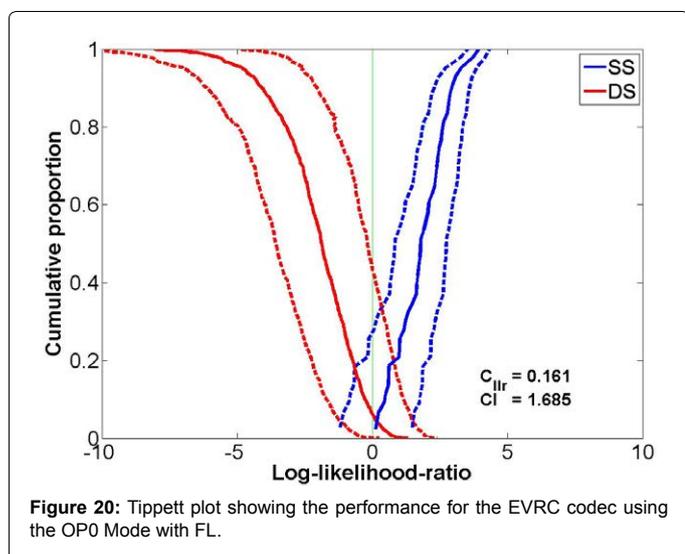
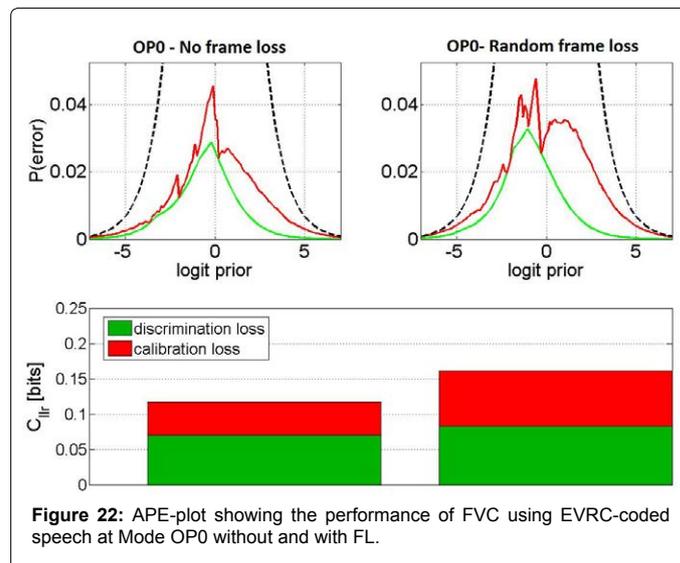
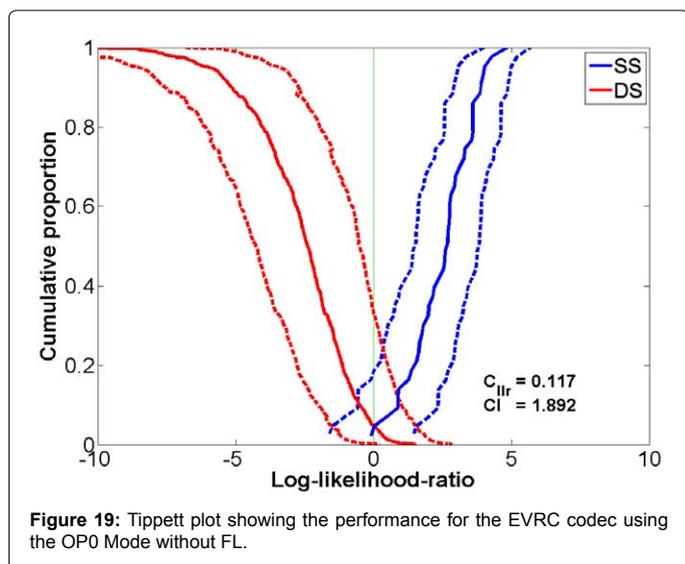
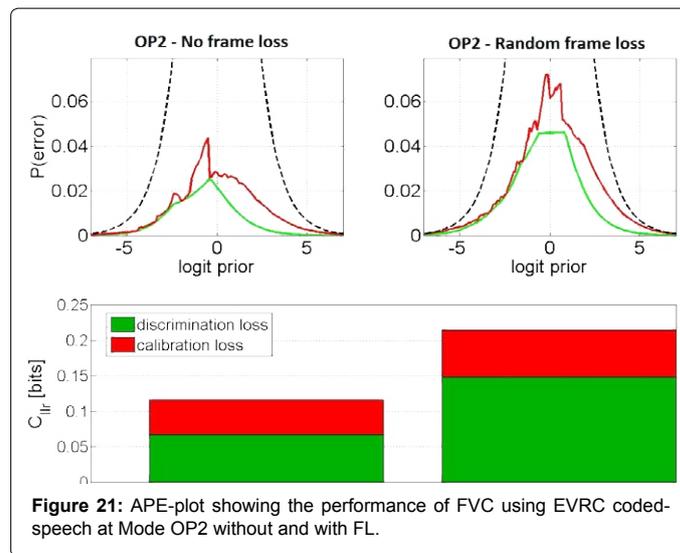
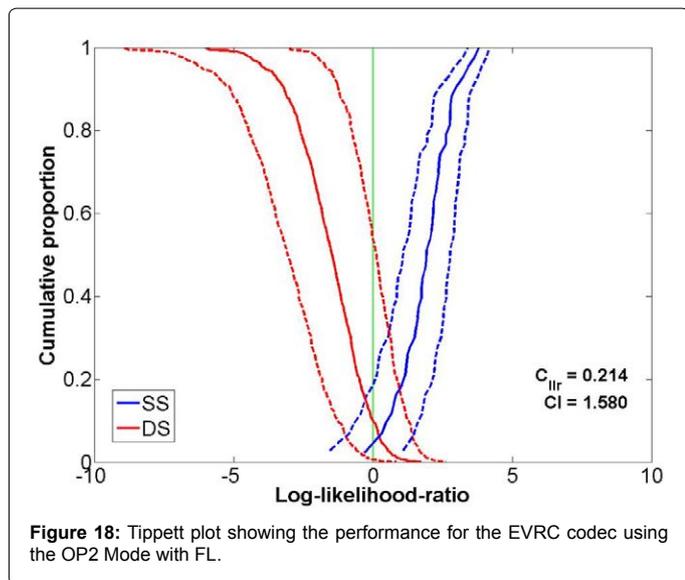
Figures 21 and 22 show APE-plots for OP2 (low quality coding)

and OP0 (high quality coding), respectively. As was the case for the AMR codec, FL in low quality coded speech causes the discrimination loss to increase significantly, in this case by about 110%. There is also a small increase in calibration loss of about 30%. The situation for high quality speech is somewhat different. Here the major impact of FL is to increase the calibration loss by about 65%, with discrimination loss increasing by only about 15%.

Conclusions

In this paper we have presented the impact of FL on FVC for speech transmitted through two major mobile phone networks: GSM and CDMA. We have noted that it is quite incorrect to assume that there is such a thing as 'generic' mobile phone speech. The GSM and CDMA mobile phone networks are fundamentally different in their design and implementation and this necessarily translates into differences in the characteristics of the speech they produce and in the subsequent impact of these differences on FVC. We have described in considerable detail the FL processes implemented by the AMR (GSM network) and EVRC (CDMA network) codecs. Our reason for describing these





processes in such detail is because we believe it essential for forensic speech scientists to have a clear appreciation of the nature and extent to which speech acquired from a mobile phone network could contain artificially generated sections. An important conclusion from this presentation is that these processes embody a considerable degree of sophistication designed specifically to mask, as far as possible, any resulting perceptual artefacts. Whether the occurrence of these processes is nonetheless still detectable from the recovered speech signal is clearly a matter for further research, but at this stage we are quite sceptical of this possibility.

We have noted that the operators of mobile phone networks permit a call to continue even if the percentage of lost frames is in the region of 10 to 15%. Given that a single lost frame will also impact upon a number of the subsequent ‘Good’ frames that follow it, the amount of artificially generated material in a mobile phone speech recording could well be higher than 10 to 15%. Our experiments have focused on vowel segments of typically 12 to 15 frames in duration. In the interests of considering worst-case conditions, we have introduced two lost frames into these segments, the temporal locations of which have been determined randomly according to a uniform distribution.

We have shown that with AMR-coded speech, FL causes a worsening of same-speaker comparisons in terms of accuracy, and noted this is more problematic for low quality coded speech than high quality. Perhaps not surprisingly, our experimental results also suggest that FL with AMR-coded speech can improve the accuracy of different-speaker comparisons. As far as reliability is concerned, FL negatively impacts upon both same- and different-speaker comparisons in a similar manner.

With the EVRC, though a number of our experimental results are similar to those of the AMR codec, there are also some important differences. One such difference is in respect to the impact of FL on the accuracy of different-speaker comparisons. For reasons which are as yet unclear, FL negatively impacts upon the accuracy of both same-speaker and different-speaker comparisons, but in terms of reliability, FL actually improves both same-speaker and different-speaker comparisons.

Though much more research needs to be done on this aspect of the impact of FL on FVC undertaken using mobile phone speech, it is clear from the results presented in this paper that it can be significant, a fact that must necessarily impact on the confidence a forensic scientist ascribes to their analysis results.

References

1. Alzqhoul EAS, Nair BB, Guillemin BJ (2015) Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison. *Science & Justice* 55: 363-374.
2. Kuhn V (1997) Applying list output Viterbi algorithms to a GSM-based mobile cellular radio system. In: *Universal Personal Communications Record, Conference Record, IEEE 6th International Conference on IEEE* pp: 878-882.
3. 3GPP, TS 45.003; 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Channel coding.
4. Morrison GS (2011) A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Comm.* 53: 242-256.
5. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Process* 10: 19-41.
6. 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.
7. Nair BB, Alzqhoul E, 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.
8. Jessen M (2014) Comparing MVKD and GMM-UMB applied to a corpus of formant-measured segmented vowels in German. In: *International Association for Forensic Phonetics and Acoustics Annual Conference (IAFPA 2014)*, Zurich, Switzerland.
9. Alzqhoul EA, Nair BB, Guillemin BJ (2014) An Alternative Approach for Investigating the Impact of Mobile Phone Technology on Speech. In: *Proceedings of the World Congress on Engineering and Computer Science*.
10. Bruhn S, Error concealment in relation to decoding of encoded acoustic signals. US. patent No. 6,665,637.
11. ETSI, Substitution and Muting of Lost Frames for Full Rate Speech Channels. Retrieved on 2 June 2013, last retrieved from <http://www.3gpp.org/>.
12. Alzqhoul EA, Nair BB, Guillemin BJ (2012) Speech Handling Mechanisms of Mobile Phone Networks and Their Potential Impact on Forensic Voice Analysis. In: *SST 2012*, Sydney, Australia.
13. 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.
14. 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.
15. 3GPP, TS 45.009 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Link adaptation.
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.
17. 3GPP2, S0018-D, Minimum Performance Specification for the Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, and 73 for Wideband Spread Spectrum Digital Systems.
18. Messer K, Matas J, Kittler J, Luettin J, Maitre G (1999) XM2VTSDB: The extended M2VTS database. In: *Second international conference on audio and video-based biometric person authentication, Citeseer* pp: 965-966.
19. Rose P (2004) *Forensic speaker identification*, CRC Press.
20. Ramos-Castro D, Gonzalez-Rodriguez J, Ortega-Garcia J (2006) Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. In: *Speaker and Language Recognition Workshop IEEE Odyssey, The IEEE* pp: 1-8.
21. D. Networks, Voice Quality Solutions for Wireless Networks. Retrieved on 21 June 2013, last retrieved from <http://www.ditechnetworks.com/>.
22. Rix AW, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)- a new method for speech quality assessment of telephone networks and codecs. In: *Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'01)*. 2001 IEEE International Conference on IEEE pp: 749-752.