Science and Justice 55 (2015) 363–374

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison

Esam A.S. Alzqhoul, Balamurali B.T. Nair, Bernard J. Guillemin *

Forensic and Biometrics Research Group (FaB)
Department of Electrical and Computer Engineering, The University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form 1 April 2015
Accepted 13 April 2015

Keywords:
GSM
CDMA
Forensic voice comparison
Likelihood ratio
Dynamic rate coding

ABSTRACT

Previous studies have shown that landline and mobile phone networks are different in their ways of handling the speech signal, and therefore in their impact on it. But the same is also true of the different networks within the mobile phone arena. There are two major mobile phone technologies currently in use today, namely the global system for mobile communications (GSM) and code division multiple access (CDMA) and these are fundamentally different in their design. For example, the quality of the coded speech in the GSM network is a function of channel quality, whereas in the CDMA network it is determined by channel capacity (i.e., the number of users sharing a cell site). This paper examines the impact on the speech signal of a key feature of these networks, namely dynamic rate coding, and its subsequent impact on the task of likelihood-ratio-based forensic voice comparison (FVC). Surprisingly, both FVC accuracy and precision are found to be better for both GSM- and CDMA-coded speech than for uncoded. Intuitively one expects FVC accuracy to increase with increasing coded speech quality. This trend is shown to occur for the CDMA network, but, surprisingly, not for the GSM network. Further, in respect to comparisons between these two networks, FVC accuracy for CDMA-coded speech is shown to be slightly better than for GSM-coded speech, particularly when the coded-speech quality is high, but in terms of FVC precision the two networks are shown to be very similar.

1. Introduction

The number of crimes committed using mobile phones has significantly increased in the last decade, with the result that mobile phone recordings are being increasingly used as evidence in courts of law. In such cases, forensic speech scientists are typically engaged to undertake an analysis of suspect and offender recordings in order to assess the strength of the evidence presented, a process often referred to as forensic voice comparison (FVC) [1–4]. But when undertaking this task it may be erroneously assumed that all mobile phone networks are similar in respect to their underlying technology, and therefore in their potential impact on the speech signal [5]. However, within this arena there are a number of network providers utilizing a variety of technologies, such as global system for mobile communications (GSM) and code division multiple access (CDMA). These two network technologies are fundamentally different in their design and internal operation and any assumption that they impact similarly on the speech signal is not correct. In respect to the geographical coverage of these networks, recent surveys have reported that the GSM network has operators in 212 countries with approximately 3 billion users [9]. In comparison, the total number of users utilizing the CDMA technology worldwide is approximately 500 million. Nonetheless, the CDMA network is still very popular in North America, China and India, with a presence in 118 countries worldwide [10,11].

The primary goal of this paper is to investigate the impact of a key feature of these mobile phone technologies, namely dynamic rate coding (DRC), on the outcome of a FVC [5]. DRC is a process of dynamically changing the source coding bit rate on a frame-by-frame basis. One key difference in respect to the GSM and CDMA networks is the mechanism driving this process. With the GSM network it is changing channel conditions, referred to as channel quality, which is the driver; with the CDMA network it is changing user demand, referred to as

Abbreviations: ACS, active codec set; ADR, average data rate; AMR, adaptive multi rate; APE, average probability of error; C/I, carrier to interference ratio; CDMA, code division multiple access; CELP, code excited linear predictive; CI, credible interval; C_{llr}, log-likelihood-ratio cost; C_{llrmin}, discrimination loss; C_{llrcal}, calibration loss; DRC, dynamic rate coding; EVRC, enhanced variable rate codec; FVC, forensic voice comparison; FDMA, frequency division multiple access; FR, full rate; GMM-UBM, Gaussian Mixture Model-Universal Background Model; GSM, global system for mobile communications; HR, half rate; ICM, initial codec mode; LA, link adaptation; LR, likelihood ratio; LLR, log-likelihood-ratio; MFCCs, Mel-frequency cepstral coefficients; MVKD, multivariate kernel density; NELP, noise excited linear prediction; NS, noise suppression; OP, operating points; PAV, pool adjacent violators; PCA, principal component analysis; PCAKLR, principle component analysis kernel likelihood ratio; PPP, pitch period prototype; TDMA, time division multiple access; UKD, univariate kernel density.

* Corresponding author at: Dept. of Electrical & Computer Engineering, The University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand. Tel.: +64 9 373 7599x88190; fax: +64 9 373 7461.

E-mail addresses: ealz002@aucklanduni.ac.nz (E.A.S. Alzqhoul), bbah005@aucklanduni.ac.nz (B.B.T. Nair), bj.guillemin@auckland.ac.nz (B.J. Guillemin).

http://dx.doi.org/10.1016/j.scijus.2015.04.006
1355-0306/

channel capacity [6,7]. The source coding bit rate directly impacts on the resulting coded-speech quality. Mobile phone networks incorporate highly sophisticated speech coding blocks, called codecs, which code the speech in order to achieve a reasonable level of data compression (i.e., low bit rate). The most widely used speech codecs in the GSM and CDMA networks are the adaptive multi rate (AMR) codec and the enhanced variable rate codec (EVRC), respectively. These codecs have many modes of operation, which in turn govern, among other aspects, the resulting bit rate per frame [8]. What the network does is initiate changes between these modes.

The approach used in this paper for evaluating the strength of evidence in FVC is via the calculation of likelihood ratios (LRs). It is acknowledged that internationally this LR framework is highly controversial and disputed [31–33]. Often referred to as the "Bayesian approach", it is regarded by some as the only logical and coherent approach to forensic science [33]. Others argue that in forensic casework that involves speech, LRs are problematic for the following reasons and should be avoided [31,32]:

[a] In order to estimate the strength of evidence with a LR, a reference or background sample from the relevant population is needed. A number of factors must be considered when choosing this including the speaker's regional and social background, environmental factors that have significant effect on voice and language, the recording situation, etc. It is argued that the range of these factors that would need to be controlled in order to establish a reliable reference sample could well be insurmountably large.
[b] In order to calculate a LR one needs information on the distribution of the speech feature of interest in the population. Aside from a few global speech characteristics such as fundamental frequency, articulation rate, and stammering, there is a lack of information on the distribution of speech features in the population.

Notwithstanding this ongoing debate, LRs are widely acknowledged in other forensic scientific disciplines. Further, even among speech forensic scientists there seems to be general agreement that they deserve merit and are inherently interesting.

Within the LR framework different methods have been established to evaluate the speech evidence, such as multivariate kernel density (MVKD) [12–14], Gaussian mixture model-universal background model (GMM-UBM) [14,15], and principle component analysis kernel likelihood ratio (PCAKLR) [16,17]. 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 has been primarily designed for token-based analysis scenarios [26]. PCAKLR has been designed to be functionally very similar to MVKD, and is thus also primarily intended for token-based analysis scenarios. MVKD and PCAKLR differ, though, in respect to the number of input speech features permitted. With MVKD this is quite small (3-4), whereas PCAKLR can handle much larger numbers of features. The experiments presented in this paper have used vowel tokens, with each being represented by 23 Mel-frequency cepstral coefficients (MFCCs). We have therefore opted to use PCAKLR for computing LRs.

The idea behind PCAKLR is simple. Firstly the set of input parameters is transformed into a new set of orthogonal (i.e., highly uncorrelated) parameters using principal component analysis (PCA). LR values are then calculated from this using univariate kernel density (UKD) analysis and their product taken to produce an overall LR based on the naive Bayesian approach [16]. Concerns might be raised that the PCAKLR approach is based on mutually contradictory suppositions, thus bringing into question the meaningfulness of any results it produces. The first of these suppositions is that the data are sufficiently multivariate normal that it is appropriate to use PCA; then that the distributions on each derived dimension are sufficiently non-normal that they should be modelled using a kernel density estimator rather than a Gaussian distribution. It is acknowledged that strictly speaking there is this contradiction. Nonetheless, in our experience the model empirically produces good results, which is our justification for using it in this investigation.

With reference to the MFCCs used in our experiments, it is known that cepstral coefficients are generally sensitive to transmission artefacts in landline networks and several compensation techniques [18–20] have been proposed to account for this, as well as a number of other factors as well. Though transmission artefacts do impact on the speech signal in mobile phone networks, the manner in which they do so is entirely different. For example, speech data is transmitted in frames. If a frame gets lost or irrecoverably corrupted during transmission, this will be detected by sophisticated error detection routines. A new frame will then be inserted for a corrupted frame using information from previous good speech frames [21,22]. As a result, partially corrupted speech data, as might occur due to channel noise, never arrives at the receiving end. Thus the compensation techniques referred to, if used to account for transmission artefacts, are not appropriate when working with mobile-coded speech [5,8].

The 23 MFCCs in our experiments have been extracted from the entire vowel segment. The maximum number of MFCCs that can be used in an analysis is determined by the sampling frequency of the speech data, this being 8 kHz, which is the standard value used in the GSM and CDMA networks. It needs to be acknowledged, though, that it is more usual in an FVC analysis to use only the first 12-14 MFCCs extracted from stationary speech frames, along with their 1st or 2nd order derivatives (i.e., deltas and delta-deltas). Preliminary experiments we have conducted seem to point to MFCCs extracted in that manner introducing higher variation in the resulting LR values when applied to mobile-coded speech [23]. In contrast, those extracted from the entire vowel segment produced less variation and, therefore, resulted in a better precision.

The speech codec in a mobile phone network is the only component that directly handles the speech signal, and therefore it is this component alone that determines the quality of the resulting transmitted speech [5,8,24,25]. Factors such as poor channel conditions, channel noise, congestion related to the number of users, etc., cannot impact directly on the speech signal, but rather indirectly by way of instructions sent to the codec from upper levels of the network to change its mode of operation to accommodate these external factors. Therefore, in order to understand comprehensively how the process of DRC might impact on the speech signal, the best strategy we believe is to fully understand all the possible modes of operation of these speech codecs and the underlying rules under which these might be initiated. We consider this to be a much better strategy than conducting a large number of experiments involving transmission of speech across an actual mobile phone network.

In our view this latter approach has two major drawbacks. Firstly, it can at best provide information on the impact on the speech signal under only a small (and unknown!) subset of transmission factors, not the totality of all possibilities. These transmission factors include: (i) DRC, namely the many bit-rate combinations that could be imposed on a sequence of speech frames (i.e., the subject of this article), (ii) frame-replacement mechanisms employed when frames are lost or corrupted during transmission, (iii) strategies implemented to lessen the impact of background noise at the transmitting end on the speech coding processes, and (iv) the impact on the coding processes of the characteristics of the microphone at the transmitting end as well as its relative placement to the speaker's mouth. It is the impact of the totality of all these possibilities that the forensic speech scientist needs to consider when drawing conclusions from their analyses. The second major drawback associated with the approach involving transmission of speech across an actual mobile phone network is the impossibility of investigating the impact of any of the above factors in isolation. If one cannot determine and analyse the impact of each factor in isolation, one will never be in a position to devise strategies for combating those impacts.