INTERACTION OF VOICE OVER INTERNET PROTOCOL
SPEECH CODERS AND DISORDERED SPEECH SAMPLES

Vijay Parsa and Donald G. Jamieson

National Centre for Audiology
University of Western Ontario, London, Ontario, Canada N6G 1H1.
parsa@nca.uwo.ca, jamieson@nca.uwo.ca

ABSTRACT

Voice over IP (VoIP) is an emerging technology where voice is transmitted over packetized data networks instead of the traditional public switched telephone networks. The effective functioning of a VoIP system depends upon the quality of coding/decoding scheme, and on the network’s performance in transmitting the voice packets. Traditional measures of speech coder quality and network performance employ voice samples from normal talkers and it is unclear what impact abnormal voice samples will have on the performance of a VoIP system. The objective of this paper is to investigate the influence of speech samples from talkers who have a form of common voice disorder on the performance of three speech coders typically used in VoIP systems. Our results show that there is a significant difference in the performance of low bit rate speech coders with disordered speech samples, and that the packet loss affects certain coders more than the others.

1. INTRODUCTION

The Internet Protocol (IP) is the protocol by which the data is transmitted from one computer to another on the public internet or private intranets and extranets. Voice over IP (VoIP) is a relatively new technology that aims to transport voice traffic over these data networks using the same protocol [1,2,3]. The VoIP systems employ speech coders that compress the speech information into fewer bits per second while maintaining the quality of input speech. Thus VoIP offers significant advantages over public switched telephone network (PSTN) in efficient utilization of the transmission bandwidth. Through the effective use of the available bandwidth, VoIP also enables the transportation of data, voice and video over the same network.

There are two key factors that will ultimately determine if the VoIP is a credible alternative to the circuit-switched phone networks - voice quality, and network reliability and performance. The voice quality provided by a VoIP system depends upon such factors as the speech coding/decoding process, end-to-end delay for voice transmission, and the loss of voice packets during transmission [1].

The speech coders that are used in the present VoIP systems are standardized by the International Telecommunications Union (ITU). The standardization process usually involves subjective listening tests such as the Mean Opinion Score (MOS) tests where the quality of code-processing speech is given an overall rating on a scale of 1 to 5. However, the standardization process includes speech samples collected only from normal talkers and these coders have not been evaluated with respect to their effects on the speech of persons with disorders of voice and/or speech. This is a significant omission, because providers of public communications systems are mandated to ensure that their services are accessible to all persons, including those with communication disorders [5,6]. In addition, studies of the impact of packet loss on voice quality for VoIP service employed only normal speech samples [7]. Although there are studies investigating the performance of speech coders with languages having characteristics very different than English [8,9], there has been no investigation into the interaction of the speech coder performance with speech disorders. In this study, the interaction between speech coders and speech disorders, and the influence of voice packet loss on the reconstruction of disordered voice samples were investigated using standardized objective measures of speech quality.

1.1. Speech coders used in VoIP systems

The G.728 [10] is a Low Delay - Code Excited Linear Prediction (LD-CELP) coder that is based on backward adaptive linear prediction and analysis by synthesis coding principles. The LD-CELP coder employs a codebook of 1024 vectors. The input speech signal is divided into frames of 5 samples and for each frame 1024 synthesized segments are generated by passing each of the codebook vectors through a gain scaling unit and a synthesis filter. The 10 bit index of the codebook vector that results in a minimum frequency-weighted mean-squared error between the original frame and the synthesized frame is transmitted. The synthesis filter coefficients and gain are updated with each frame in a backward adaptive manner using the information from the previously quantized frames. The LD-CELP coder circumvents the pitch estimation procedure by employing a longer (50th order) synthesis filter.

The G.729 coder is an example of a forward linear prediction, analysis by synthesis coder. Here the input speech is divided into frames of 10 ms. For each frame a 10th order linear prediction filter is derived through the Levinson-Durbin algorithm [11]. The 10 LP coefficients are transformed into Line Spectrum Pairs (LSPs) and are quantized using a two stage predictive vector quantization. Two codebooks are used to specify the excitation signal parameters: an adaptive codebook for determining the pitch delay and gain, and a fixed codebook for estimating the residual signal and gain.

The G.729 coder is an example of a forward linear prediction, analysis by synthesis coder. Here the input speech is divided into frames of 10 ms. For each frame a 10th order linear prediction filter is derived through the Levinson-Durbin algorithm [11]. The 10 LP coefficients are transformed into Line Spectrum Pairs (LSPs) and are quantized using a two stage predictive vector quantization. Two codebooks are used to specify the excitation signal parameters: an adaptive codebook for determining the pitch delay and gain, and a fixed codebook for estimating the residual signal and gain.
parameters. These parameters are computed for subframes of 5 ms each. This information is encoded into 80 bits per 10 ms frame resulting in an overall rate of 8 kbps.

The G.723.1 is a dual rate speech coder operating at 6.3 kbps or 5.3 kbps, with the higher bit rate providing better quality speech. At the encoder, the input speech is divided into 30 ms frames which are in turn subdivided into subframes of 7.5 ms [12]. A 10th order LP filter is calculated for each of the subframes which is used in the construction of the perceptual weighting filter. The LP coefficients from the last subframes are quantized using predictive split vector quantization. The open loop pitch period is determined every two subframes, and is refined by a closed-loop pitch predictor operating on each of the subframes. The non-periodic component of the excitation signal is generated by Multi-Pulse Maximum Likelihood Quantization (MP-MLQ) for the higher bit rate coder, or through an algebraic codebook search for the lower bit rate coder.

1.2. Speech quality measures

Objective measures of speech coder performance attempt to compute the “distance” between the coder-processed speech and the original speech. Two recently standardized measures are used in this manuscript, the Measuring Normalizing Blocks (MNB) measure [13], and the Perceptual Speech Quality Measure (PSQM) [14].

Within PSQM, the coder-processed and original speech samples are mapped into psychophysical representations that match the internal impression of the speech samples as closely as possible. These internal representations make use of psychophysical equivalents of frequency, intensity and masking. Finally, the Noise Disturbance (ND), which represents the difference in the internal representations of the processed and original speech samples, is computed as a measure of the coder quality.

Recently, the MNB technique has been shown to better the performance of the PSQM [13]. The MNB measure also involves a perceptual transformation in frequency (Hertz - to - Bark scale transformation), and intensity (logarithmic transformation). In addition, normalization on different frequency and time scales is performed in an attempt to match the internal representation of the speech sample. Finally an Auditory Distance (AD) parameter is calculated as the weighted distance between the perceptually transformed processed and original speech samples.

2. METHOD

2.1. Speech samples

All speech samples were selected from the voice disorders database recorded at the Massachusetts Eye and Ear Infirmary [16]. The recorded samples from thirty patients representing a range of organic, neurological, traumatic, and psychogenic voice disorders and severities were selected for use in the study. Samples were five seconds of the Rainbow passage recorded in a sound-proof booth at the Massachusetts Eye and Ear Infirmary. For comparison purposes, additional speech samples were obtained from twenty speakers having normal hearing and speech, and English as their first language. The normal talkers exhibited no abnormal vocal characteristics and indicated no history of voice disorders.

Each of the original data files was down-sampled to 8 kHz, and processed through each of the four coders, yielding five versions of each speech sample for evaluation: 1) the original, unprocessed file, 2) a G.728-coded file, 3) a G.729-coded file, 4) a G.723.1-coded file at 6.3 kbps, and 5) a G.723.1-coded file at 5.3 kbps.

In addition, packet loss conditions were simulated by randomly discarding a specified number of frames from the coded speech information. At the decoder, the information from the frame that preceded the discarded frame was used to reconstruct the segment corresponding to the discarded frame. Conditions resulting in 1, 5, 10 and 20 percent packet loss were simulated and the objective speech quality measures were calculated for each of the packet loss conditions for all the coders.

2.2. Objective quality measurement

The coder-processed and original speech signals were first time-aligned using the cross-correlation procedure. The mean values were then subtracted and the RMS levels were normalized respectively.

The MNB measure was implemented following the procedure outlined in Voran [13]. The coder-processed and original speech samples were divided into frames of 128 samples, with a frame overlap of 50%. Each of these frames was multiplied by the Hamming window and transformed into the frequency domain using the Fast Fourier Transform (FFT). A Frequency Measuring Normalizing Block (FMNB) was applied to cover all the frames of the input signals and four measurements, which cover the lower and upper band edges of the telephone band, were saved. Eight additional measures were computed from MNB Structure 1. The Auditory Distance (AD) measure was a linear combination of these twelve measures. The AD value is put through a logistic function, the parameters of which were optimized empirically, to give the L(AD) parameter as the final measure of coder-processed speech quality.

The PSQM was computed according to the guidelines provided in the ITU P.861 document. The processed and original speech waveforms were divided into frames of 256 samples, with a frame overlap of 50%. The sampled spectral density of each frame was then computed by squaring the magnitudes of frame Fourier transforms. The spectra were then frequency-warped by mapping the Hertz scale to the critical band scale, and processed through filters simulating the telephone receiving characteristics and Hoth noise. Sampled compressed loudness densities were then computed by warping the intensity scale using Zwicker’s compression function. Finally, a cognitive model was applied which takes into effect the loudness scaling, internal cognitive noise, and asymmetry processing parameters and a Noise Disturbance (ND) was computed.

3. RESULTS

The mean and standard deviation of the objective speech quality parameters for all four coders for the normal and disordered talker groups are reported in Table 1. Also shown in this table are the results of independent samples t test for
testing the equality of means for normal and pathological talker groups. As expected, both objective measures show a degradation in the quality of reconstructed speech as the bit rate decreases. More importantly, there is a significant difference between the objective measures for the normal and disordered talker groups for all the coders except the G.728 coder. The reconstructed versions of pathological voice samples were further away from their original, unprocessed counterparts for G.729, and G.723.1 coders in comparison with normal speech samples processed by the same coders. With the exception of G.728 coder, the standard deviation of the objective quality measures is also higher for the pathological talker group, indicating that quality was more variable for coder-processed disordered speech than for normal speech group.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Group</th>
<th>G.728 (0.036)</th>
<th>G.729 (0.075)</th>
<th>G.723.1(6.3) (0.061)</th>
<th>G.723.1(5.3) (0.074)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norm</td>
<td>0.922</td>
<td>0.905</td>
<td>0.891</td>
<td>0.870</td>
<td></td>
</tr>
<tr>
<td>MNB</td>
<td>0.897</td>
<td>0.788</td>
<td>0.786</td>
<td>0.760</td>
<td></td>
</tr>
<tr>
<td>Path</td>
<td>2.425</td>
<td>4.134</td>
<td>3.769</td>
<td>3.472</td>
<td></td>
</tr>
<tr>
<td>$t$</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>Norm</td>
<td>1.100 (0.172)</td>
<td>1.617 (0.275)</td>
<td>1.591 (0.311)</td>
<td>1.823 (0.375)</td>
<td></td>
</tr>
<tr>
<td>Path</td>
<td>1.113 (0.145)</td>
<td>2.225 (0.405)</td>
<td>2.054 (0.491)</td>
<td>2.295 (0.502)</td>
<td></td>
</tr>
<tr>
<td>$t$</td>
<td>-0.287</td>
<td>-5.863</td>
<td>-3.739</td>
<td>-3.587</td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>0.78</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td>p&lt;0.05</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Mean L(AD) and ND values for normal and abnormal talker groups along with the $t$ values and their significance.

Figure 1 depicts the effect of packet loss on the quality of reconstructed speech. The G.728 coder exhibited a higher degree of sensitivity to packet loss. The primary reason for this phenomenon is the backward adaptive nature of the G.728 coder where previous frame information is used to update the linear prediction parameters used in the resynthesis of the input speech. Thus loss of frames corrupts the parameter update process which results in lower quality reconstructed speech. With no packet loss, the performance of the G.728 was uninfluenced by the type of input speech sample. However, even with 1% packet loss there appears to be differences between normal and pathological talker groups. The performance of the G.729 coder was also affected by the packet loss, and the difference in performance for normal and disordered speech samples increased at higher packet loss rates. For the G.723.1 coder operating at 6.3 kbps, the packet loss affected the reconstruction process of normal speech samples more than that of the disordered speech samples. Beyond a packet loss rate of 15%, there was no significant difference in the way the coder performed with normal or disordered speech samples. Similar observations can be made for the G.723.1 coder operating at a lower bit rate of 5.3 kbps.

4. DISCUSSION AND CONCLUSIONS

Voice over IP (VoIP) is a recent technological development in which voice signals are transported over packetized data networks instead of circuit switched networks. The Quality of Service (QoS) from a VoIP system depends upon several factors ranging from the transmission delay and jitter over the packetized network to the quality of the speech compression/decompression process. The objective of this paper was to quantify the performance of four commonly used VoIP coders, viz. the G.728 coder at 16 kbps, the G.729 coder at 8 kbps, and the G.723.1 coder at 6.3 kbps and 5.3 kbps, with speech data collected from disordered talkers.

In this report, two objective measures of speech quality, which have both been standardized by the ITU and have been shown to predict perceptual judgments of speech coder quality, were used to evaluate the performance of a speech coder. Continuous speech and sustained vowel tokens collected from 20 normal and 30 disordered talkers with a variety of voice pathologies were employed evaluate coder performance. In addition, conditions including packet loss during network transport were simulated and their influence on the quality of the reconstructed speech was evaluated.

The G.728 coder is a backward adaptive linear prediction coder that employs shorter frame sizes and longer analysis/synthesis filters. Only the index of the codebook vector that provides the best synthesized speech is transmitted for each frame. Our results show that this approach is robust to the type of input speech sample: coder reconstructed speech samples from both normal and disordered talkers were of similar quality. However, the performance of this coder is highly sensitive to packet loss conditions. With as little as 1% packet loss, there was a significant drop in the quality of the reconstructed speech irrespective of the type of the input speech sample.

The G.729 coder is a forward adaptive linear prediction coder that employs an adaptive codebook for quantizing the pitch information and an algebraic codebook for encoding the residual signal information. While this procedure provides toll quality speech with normal speech samples, its performance was affected by the pathological speech data. In addition, the disparity in the performance of the G.729 coder with normal and disordered speech samples increased with the advent of packet loss.

The G.723.1 coder is another example of a forward adaptive linear prediction coder which can operate at two bit rates : 6.3 kbps and 5.3 kbps. At both these rates, the coder transmits the LP coefficient information and pitch information through an adaptive codebook search. At 6.3 kbps, the residual signal is coded using the multi pulse - maximum likelihood procedure, while the 5.3 kbps method uses the algebraic codebook. Both these coders were found to differentially degrade the quality of input speech. These coders were more robust to the packet loss conditions, especially when pathological voice samples were used.

The difference in speech quality measures between normal and disordered talker groups appears to stem from the way the residual signal is encoded. When there was no packet loss during transmission, there was enough resolution in the G.728 codebook to resynthesize the pathological voice signals successfully at the decoder. In the G.729, and G.723.1 coders, however, an algebraic codebook structure is used in which the residual signal is quantized using a series of pulses. Our results confirm that this procedure does provide high quality reconstructed speech when samples from normal talkers were
employed. However, the inherent “noisiness” in the speech samples collected from disordered talkers results in a richer linear prediction residual. Our results show that this increased information in the residual signal was not adequately characterized by the G.729 and G.723.1 coding algorithms.

5. ACKNOWLEDGEMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC), the Ontario Rehabilitation Technology Consortium (ORTC), the Canadian Foundation for Innovation (CFI), and the Ontario Innovation Trust (OIT).

6. REFERENCES


