A Comparison of Speech Coding Algorithms

ADPCM \textit{vs} CELP

Shannon Wichman
Department of Electrical Engineering
The University of Texas at Dallas
Fall 1999
Abstract

Factors serving as constraints in today’s wireless communication systems include bandwidth and power. The desire to communicate as much information by transmitting a minimal set of data drives an entire industry. In wireless systems that require the transmission of speech, these goals are addressed by developing efficient methods of reducing the amount of information required to transmit and receive quality speech. For this reason, speech coding has been, and remains, the topic of aggressive research. This paper compares two classes of speech coding algorithms differentiated by their fundamental approach to coding and decoding speech signals. ADPCM, adaptive differential pulse code modulation, continues to extrapolate and optimize the brute force methods of direct digital signal coding. CELP, code excited linear predictive, is a general class of algorithms which attempt to model speech. These coding techniques will be compared on a qualitative and quantitative basis, discussing the fundamentals behind each and exploring factors important to both system implementers and subscribers.
Introduction
The recent exponential growth of wireless telecommunications today drives all aspects of wireless technology to higher degrees than ever before. As the Information Age transitions from the medium of physical wires to the airwaves, economical, physical, and practical barriers drive many areas of research which are all ultimately intertwined in a single objective: to transfer maximal amount of information consuming minimal relevant resources. Due to the impact on parameters such as bandwidth requirements and conversation quality, it can be argued that the most important component of any telephony system is that which generates the digital representation of the speech [1]. This project compares two of these speech coding techniques commonly employed in today’s applications: ADPCM and CELP. Adaptive differential pulse code modulation, ADPCM, and code excited linear prediction, CELP, make an interesting pair because they address speech compression using two fundamentally different techniques. After some necessary background information is covered, the technique associated with each algorithm is discussed. The goal of this paper is to demonstrate the differences of these two unique algorithms, reinforcing the ideas with quantitative and qualitative data.

Background
In the most basic form, speech coding is simply the digital representation of the analog speech signal. The brute force method to preserve quality of speech typically requires a 12 or 13 bit linear quantization. The two coding techniques which form the basis for most speech coding algorithms today, A-law and \( u \)-law pulse code modulation (PCM), retain the same quality represented with only 8 bits using logarithmic quantization [2]. These techniques typically form the parametric baseline used for all other comparisons, especially the data rate. At a sampling rate of 8 kHz, the data rate for both A-law and \( u \)-law is 64 kbps, and many algorithms today assume the input for the algorithm is already in this format. Compression is improved by exploiting properties of speech signals, which are bandwidth limited and relatively coherent from one sample to the next. Both ADPCM and CELP leverage from these characteristics.
There are essentially two approaches to speech coding: waveform coding and vocoding. ADPCM falls into the category of waveform coding, which attempts to preserve the waveform of the original speech signal. Speech coders based on CELP are vocoders, a technique that attempts to model the generation of speech [1]. The quality of waveform coders can be compared by direct comparison between the decoded waveform and the original waveform. Since vocoders do not attempt to preserve the waveform, this approach for quality comparison is not applicable. For this reason, the quality of most speech coding techniques is compared by a measure of user perception. The mean opinion score, MOS, is a number between 0 and 5 which represents the average quality as judged by listeners, 0 being the worst. A score of 4 is labeled ‘toll quality’ and is the quality of signal perceived by users in a typical long-distance call [2].

Two other characteristics often used to compare coding schemes are complexity and delay. Complexity, commonly measured in MIPS often relates to system implementation requirements. Delay is the time required for an input signal to be coded and reconstructed at the receiver and is important due to its impact on conversation quality.

**ADPCM**

ADPCM, adaptive differential pulse code modulation, is one of the most widely used speech compression techniques used today due to its high speech quality, low delay, and moderate complexity [2]. ADPCM is a waveform coder, and achieves its compression improvements by taking advantage of the high correlation exhibited by successive speech samples. Referring to Figure 1 below, the ADPCM encoder first calculates the difference between the input signal, typically A-law or \( u \)-law, and a signal generated by a linear adaptive predictor. Due to the high correlation between samples, the difference signal has a much lower dynamic range than the input signal. The reduced dynamic range means fewer bits are required to accurately represent the difference signal as compared to the input. Transmitting only the quantized difference signal reduces the transmitted data rate. A four-bit quantization translates to a 32 kbps data rate for a compression ratio of 2:1. Higher quality is achieved through the adaptive nature of the quantization.
Analyzing the time varying characteristics of the difference signal, the size of the quantization steps, and the rate at which quantization steps change facilitate higher accuracy for a wider dynamic range [2]. As mentioned, signal received at the decoder is the quantized difference signal. The ADPCM decoder is essentially the reverse process of the encoder.

More than one standard exists for ADPCM. The International Telegraph and Telephone Consultative Committee, CCITT first adopted recommendation G.721 that defines standard ADPCM operating at 32 kbps using 4 bit quantization. This standard was eventually encompassed by recommendation G.726, which expanded the original definition to include data rates of 40, 32, 24, and 16 kbps. This relates to 5, 4, 3, and 2-bit quantization of the difference signal [4]. A third recommendation was also adopted, G.727, addressing 5, 4, 3, and 2 bit sample embedded ADPCM which is effectively a variable rate definition of G.726 [5].

**Figure 1. ADPCM Encoder**

As most cellular applications demand a speech coder that can operate at under 8 kbps, even at 2-bit quantization ADPCM is considered to have high bandwidth demands. The relatively low complexity of ADPCM makes it attractive for systems, which require waveform preservation or have relaxed bandwidth
restrictions. The latter of these fits cordless telephony and microcell applications well. The G.721 ADPCM has been specified in a few popular cordless standards. Great Britain introduced the CT2 system to address second generation cordless applications [6]. The DECT, Digital European Cordless Telephony, system also covers second generation cordless, facilitating compatible systems across Europe [6]. The third generation Personal Access Communication System, PACS, also specifies voice coding using 32 kbps ADPCM [6].

CELP

In order to maintain acceptable quality below the 8 kbps data rate, a fundamentally different approach to speech coding and sizeable jump in complexity is encountered. The most widely documented scheme for speech coding operating at under 8 kbps is CELP, code excited linear prediction [3]. Cited as early as 1982 [3], CELP actually represents a class of vocoders, all computationally expensive, which are based on the analysis and synthesis of incoming speech. The scope of this project limits the focus on one variation of CELP, and for the purpose of example and demonstration, this project focuses on the U.S. Federal Standard 1016 that defines a CELP algorithm operating at 4.8 kbps.

Opposed to waveform encoding, the goal of CELP algorithms is to code and transmit the minimal amount of information necessary to synthesize speech which is audibly perceived to be accurate. This is achieved by modeling aspects of the speech generation process, effectively modeling the vocal tract [1]. Two components are generally found in these models, one which models the long term speech structure, the pitch, and another, which models the short-term speech structure, the formant. Relying on human perception as the measure of quality makes measurements of SNR irrelevant, and the overall rating of a specific implementation is in terms of mean opinion score, or MOS. The technique of incorporating synthesis as a fundamental part of signal analysis is common to CELP algorithms and is why they are often referred to as analysis-by-synthesis techniques. Another departure from ADPCM, CELP algorithms work on frames of
sampled input speech. The length and duration of the frame depends on the specific CELP algorithm. USFS-1016 operates on 30 ms frames which are further split into four 7.5 ms subframes [2]. Termined vector quantization, the frames of sampled input speech, input vectors, are compared to an existing set of excitation vectors called the code book. The search for the code book entry which produces the closest match is often an exhaustive search, making this class of algorithms computationally expensive. Each code book entry is applied to an LPC filter to generate a synthesized speech vector. The synthesized sample is compared to with the input vector and rated with a perceptual weighing filter. The index of the code book with the vector producing the closest perceptual match to the input vector is selected and transmitted. Low bit rates are achieved because only the code book indices, LPC filter coefficients, and gain information are transmitted. Figure 2 below, adapted from [2], shows the block diagram for the USFS-1016 CELP encoder. Two code books and a 10th order LPC filter are used in USFS-1016. The 256 entry adaptive code book models the long-term speech structure, and the 10th order LPC filter models the short term speech structure. The residual signal remaining after adaptive code book quantization and LPC synthesis is modeled by the static 512 entry stochastic code book [2]. The encoder transmits the index and gain for both stochastic and adaptive code books each subframe, or four times per 30 ms frame. The LPC filter coefficients are transmitted only once per frame. Modeling the long-term speech by the 256 entry adaptive code book improves quality at the cost of additional computational load of the same complexity as the 512 entry stochastic code book look-up. At the expense of quality, bit rates below 4.8 kbps can be achieved by limiting the search of the stochastic code book.

Disadvantages associated with USFS-1016 CELP include the less than toll quality speech (3.2 MOS), and high processing delays (90 ms). The frame-by-frame analysis, coupled with the high processing demands introduces delay which can degrade quality of conversation and introduce difficulties in related speech processing components, such as echo canceling. Other standards have been introduced which addressed both of these issues. Important standards often referenced in current literature include CCITT
recommendation G.728 for LD-CELP, or low delay CELP, and G.729 for CS-ACELP, conjugate structure algebraic CELP. Both offer quality improvements but have higher bit rates (16 kbps and 8 kbps respectively). Less complex than LD-CELP, CS-ACELP operates at half the bit rate and offers toll quality speech [2]. Yet another algorithm, VSELP, is defined in the IS-54 standard [6].

![Figure 2. USFS-1016 CELP Encoder](image)

**Results**

The results presented here are the products of samples processed by software implementations of G.721 [7] and USFS-1016 [8]. Since speech widely varies between age, gender, and ethnicity, etc., speech encoders must adequately support a wide range of speech characteristics. Tables 1 and 2 below present data and processed samples for typical male and female speech respectively. The quantitative data is extracted from the processed sample sizes and data presented in [2]. Figure 3 below demonstrates the difference between waveform coding and vocoding. The samples are fragments from the male speech sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample Size</th>
<th>Bit Rate</th>
<th>Quality</th>
<th>Ratio</th>
<th>MIPS</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>27585 bytes</td>
<td>64 kbps</td>
<td>4.3 MOS</td>
<td>1.0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G.721 ADPCM</td>
<td>13793 bytes</td>
<td>32 kbps</td>
<td>4.1 MOS</td>
<td>2.0</td>
<td>10-16</td>
<td></td>
</tr>
<tr>
<td>USFS-1016 CELP</td>
<td>4218 bytes</td>
<td>4.8 kbps</td>
<td>3.2 MOS</td>
<td>6.5</td>
<td>13-25</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Demonstration of Male Speech.
Table 2. Demonstration of Female Speech

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sample Size</th>
<th>Bit Rate</th>
<th>Quality</th>
<th>Ratio</th>
<th>MIPS</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Sample</td>
<td>45001 bytes</td>
<td>64 kbps</td>
<td>4.3 MOS</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G.721 ADPCM</td>
<td>22501 bytes</td>
<td>32 kbps</td>
<td>4.1 MOS</td>
<td>2.0</td>
<td>10-16</td>
<td></td>
</tr>
<tr>
<td>USFS-1016 CELP</td>
<td>6919 bytes</td>
<td>4.8 kbps</td>
<td>3.2 MOS</td>
<td>6.5</td>
<td>13-25</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Waveforms resulting from ADPCM and CELP encode & decode.

It is evident, from visual comparison, the difference between waveform coding and vocoding. The mean square error is calculated for each algorithm against the original signal. Note that the standard deviation if the original speech sample is 0.0747.
Conclusion

The intention of this project was to compare two fundamentally different speech coding techniques, ADPCM and CELP. ADPCM, a waveform coder, minimizes the number of bits by coding the speech in the form of a difference signal that requires fewer quantization bits to represent. At 4-bit quantization, ADPCM operates at 32 kbps, which is still considered relatively high. The low computational requirements make it an attractive solution for applications with relaxed bandwidth constraints, such as indoor and microcell use. The low MSE error of produced by the ADPCM example demonstrates effectiveness of the waveform coding technique. CELP algorithms model speech, transmitting only parameters necessary to synthesize speech that is perceived to sound like the original. Bit rates as low as 4.8 kbps can be achieved by transmitting only the parameters which represent the model. Toll quality CELP algorithms exist that require additional computational power. Still, the low bit rates offered make them attractive for cellular applications. The large MSE produced by the CELP algorithm, with relationship to the quality of speech produced, demonstrates the effectiveness of the CELP technique of modeling speech.

References