EXCITATION CODEBOOK DESIGN FOR CODING OF THE SINGING VOICE

Youngmoo E. Kim

MIT Media Lab
Machine Listening Group
20 Ames St., E15-401, Cambridge, MA 02139
moo@media.mit.edu

ABSTRACT

The technique of Code Excited Linear Prediction (CELP) has led to the development of voice coding systems that provide toll quality speech at very low bitrates. While speech and singing share many similarities in terms of production, standard speech coding implementations fall far short when transmitting the singing voice. This paper explores the reasons for this discrepancy and suggests new variations on CELP speech coders that specifically enhance the quality of encoded singing for individual singers. These modifications could be used in a low-bitrate singing voice codec which, in conjunction with multi-track structured coding schemes such as MPEG-4 Structured Audio, could provide a highly compressed yet high-quality representation of a complex audio scene.

1. INTRODUCTION

A great deal of research has gone into voice coding, though most of it has focused on efficient coding of speech. The most common implementation in use today is Code-Excited Linear Prediction (CELP), which describes a family of coding techniques based on Linear Predictive Coding (LPC) for the calculation of filter coefficients, roughly corresponding to the resonant (wideband) response of the vocal tract, along with modeling of the excitation vector, roughly corresponding to the glottal excitation function produced by the vocal folds. Different CELP implementations model the excitation in slightly different ways, but all use a codebook or codebooks of predefined excitation vectors, with the codebook(s) stored at both the transmitter and receiver. For each frame, a vector in the codebook is selected that results in the best speech reconstruction according to some objective criteria. Since both transmitter and receiver have the codebook(s), all that needs to be transmitted is a codebook index. In this way the coding parameters are kept to a minimum and the overall bitrate is greatly reduced.

Although both are produced using the same mechanisms, perceptually important qualities of the singing voice are quite different from that of speech. These differences make speech codecs less applicable for high-quality transmission of singing. This research investigates these differences and explores ways of modifying a standard CELP coder to better transmit the singing voice. Specifically, data taken from singing as opposed to speech is used to create codebooks that are more applicable for singing. The resulting codec removes some of the artifacts generated when using CELP encoding for singing, providing toll quality transmission of singing while maintaining an overall low bitrate.

The long-term goal of this ongoing research is the development of models that can be used for parametric synthesis of the singing voice. Previous work [1] focused on techniques for modeling the resonant characteristics of individuals' singing voices for more efficient transmission with an eye towards parametric manipulation. Likewise, this paper focuses on the modeling of the excitation function for individual singers for more efficient transmission, but again with the ultimate goal of parameterization.

2. BACKGROUND

This section provides background material for the research that follows in subsequent sections, particularly issues related to voice coding, the ways in which singing differs from speech, the context provided by structured audio, and the CELP algorithm.

2.1. Vocal Production

Artificial generation of the human voice, speaking or singing, continues to be an elusive goal. Acoustically, this is because of the dynamic variation involved in the vocal process. Different words require different positions of jaw, tongue, teeth, etc., changing the configuration and therefore the resonant properties of the vocal tract. Vocal sounds begin with the breath pressure from the lungs, which forces open the vocal folds. The folds are then pulled back together by Bernoulli force. Repetition of this process results in a periodic excitation, which correlates to our perception of pitch (fundamental frequency). Reconfiguration of the shape of the vocal tract (throat, mouth, nose, tongue, teeth, and lips) for different syllables creates different filters, and the filtered output from the excitation is perceived as different phonemes. Vocal sounds are characterized as voiced or unvoiced, depending on whether phonation occurs for the sound. All vowels and some consonants (e.g. [m], [n], [l]) are voiced, while other consonants (e.g. [f], [s], [t]) are not. For unvoiced sounds,

the excitation is no longer the phonation of the vocal folds, but the turbulence caused by air impeded by the vocal tract, resulting in a noise-like excitation. Some consonants (e.g. [v], [z]) are mixed sounds that use both phonation and turbulence to create the overall sound [2].

Speech and singing research have always been closely related, but there are significant differences between the two. In English, speech consists of approximately 60% voiced sounds and 40% unvoiced sounds, while the vast majority of sounds generated during singing are voiced (>90%) [3]. In singing, each note that is sung is fairly constant and quantized in pitch (in Western music), as opposed to speech, in which pitch varies unpredictably and continuously. In the most common classical singing technique, known as *bel canto*, singers are taught that vowel sounds (consisting of several resonant peaks) are the most efficient sounds for singing and should be held as long as possible between consonants. Singers also learn to develop a high degree of consistency in the pronunciation of vowels, making it easier to determine the vowel from analysis of the signal [1]. These distinctions must be taken into account when designing efficient model-based encoding schemes for singing.

2.2. Structured Audio

Structured audio refers to research on the creation, transmission, and rendering of parametric sound representations [4]. Parametric, low-dimensional models exist for simulating many musical instrument sounds. One standardized foundation for such algorithmic representations is MPEG-4 Structured Audio [5]. Currently, there is no such low-dimensional high-quality model for the singing voice. An ideal structured singing voice model would be able to use musical knowledge on a symbolic level (the score, lyrics, etc.) for very compact representation. The parameters in such a model would also be modifiable, allowing for better manipulation and interaction. Such a model is the ultimate goal of this research.

The MPEG-4 Audio standard also allows for *scene description*, which provides structure at a higher level by allowing the separation of encoding for different sound sources. With *a priori* knowledge of the source type, specific encoding schemes can be used for maximum compression. For example, peech tracks can be transmitted using a speech codec, such as CELP; acoustic instruments can be encoded using one of the natural audio codecs; and synthetic sounds can be represented using MPEG-4 Structured Audio. It is possible for some complex sound mixtures to be encoded at high-quality at a greatly reduced bitrate. In this context, the applications of an encoder specifically for singing voice become obvious (Figure 1).

MPEG-4 Structured Audio can also be used to implement arbitrary audio coding techniques, including perceptual transform coding and CELP. This technique is called *generalized audio coding* [6], and can lead to hybrid coding techniques combining aspects of traditional audio coders with the flexibility of synthesis [1]. It also facilitates the rapid development and deployment of new codecs, so that a new codec (e.g. for an specific individual's singing voice) could be downloaded along with a piece of content encoded using that codec, removing dependence on fixed hardware implementations of sound encoding and decoding.

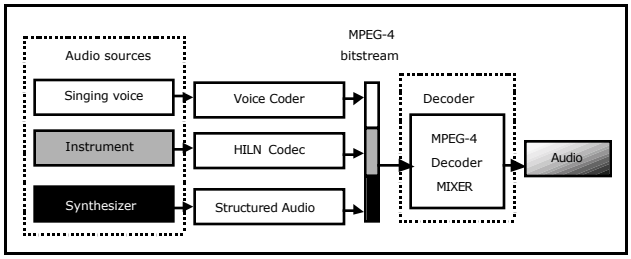

Figure 1: *Multiple source encoding using MPEG-4 Scene Description*

2.3. The CELP algorithm

The CELP codec [7] has proven to be quite successful at transmitting toll-quality speech at low-bitrates (down to 4 kbits/sec) and intelligible speech at even lower bitrates. Its use is widespread in today's digital communications devices, such as cellular phones. CELP is a combination of traditional Linear Predictive (LP) modeling of the resonant qualities of the vocal tract coupled with complex excitation modeling combining deterministic and stochastic sources.

Linear predictive analysis calculates an estimate, $\tilde{s}[n]$ to the source signal $s[n]$, using a linear combination of p past samples of the input signal:

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{1}$$

The transfer function relating the source signal and the signal estimate is easily shown [8] to be an all-pole filter:

$$H[z] = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{2}$$

This equation shows linear predictive analysis to be equivalent to a source-filter model, where the vocal tract response is modeled using a time-varying all-pole filter function of order p. Common values for p range from 10 to 20. The LP parameters are calculated using the well-known autocorrelation or covariance techniques (see [9] for details).

The excitation is usually modeled on a finer scale than the LP analysis (several subframes per LP frame). The general method used in the selection of the excitation is a closed-loop analysis by synthesis, which occurs in two parts. The first part is a pitch analysis from a buffer of recently generated excitations. The assumption is that the excitation pulses in voiced sounds are fairly regular, and an excitation from a previous period can approximate the excitation of the current period. A linear search is performed on the buffer over the range of expected fundamental periods, and the vector that best synthesizes (being filtered with the LP parameters) the current waveform frame is selected. Sometimes this search is aided using a low-order (usually n≤3) pitch prediction filter. Also, in every other frame the search range is sometimes limited to

be in the neighborhood of the previously determined pitch period, reducing complexity and bandwidth (since the period can be transmitted in fewer bits). After this vector is chosen, a gain calculation is performed to further optimize the match between the input waveform and the synthesized output. This process of excitation generation is sometimes referred to as an *adaptive codebook* or *self-excitation*.

The second part of the excitation generation consists of a search through a static codebook of stochastic vectors to model unvoiced excitations. The residual error from the previous pitched excitation determination is compared against the entries of the codebook, filtered by the LP parameters. The entry corresponding to the resulting waveform that best matches (determined via correlation) the residual is chosen, along with a corresponding gain. Since the transmitter and receiver share a copy of the codebook, the codebook entry number and gain are the only values need be transmitted for this part of the excitation. Codebooks vary according to implementation, but generally contain on the order of hundreds of entries. The entries themselves may be completely random, since they are intended to generate the noise-like excitations of unvoiced speech.

No voiced/unvoiced decision of the input signal frames is required. The excitation will be composed of some ratio of the deterministic and stochastic vectors, which is established by the gain parameters calculated.

3. SINGING VOICE CODEBOOK DESIGN

The following section explores ways of modifying the CELP codec to better serve in transmitting the singing voice. The deficiencies in applying CELP to singing voice are identified and analyzed.

3.1. CELP and the Singing Voice

Since singing is comprised primarily of voiced sounds, we would expect that the modeling of the voiced excitation using CELP would be fairly successful for singing. This is generally the case, though the pitch track sometime chooses multiples of the fundamental period. For the most part, this is tolerable since excitations in singing tend to be quite consistent. Improvements could be made in this area, but greater gains are easier found elsewhere.

A larger problem stems from the fact that standard CELP implementations are designed with a dual-excitation model, one for voiced and one for unvoiced sounds. As noted above, while speech contains a fairly even mixture of voiced and unvoiced sounds, singing is almost entirely voiced. Hence, the CELP excitation model is not ideally suited for the singing voice. A comparison of the residuals after the determination of pitched and stochastic excitations for a singing voice signal reveals that the stochastic codebook does little to reduce the residual error in singing (Figure 2). As can be seen, the residual error from the singing example is hardly reduced at all after application of the stochastic excitation.
These deficiencies are clearly audible in the quality of singing encoded using CELP. This is readily experienced by attempting to sing over digital cellular phones, which use variations of CELP coding.

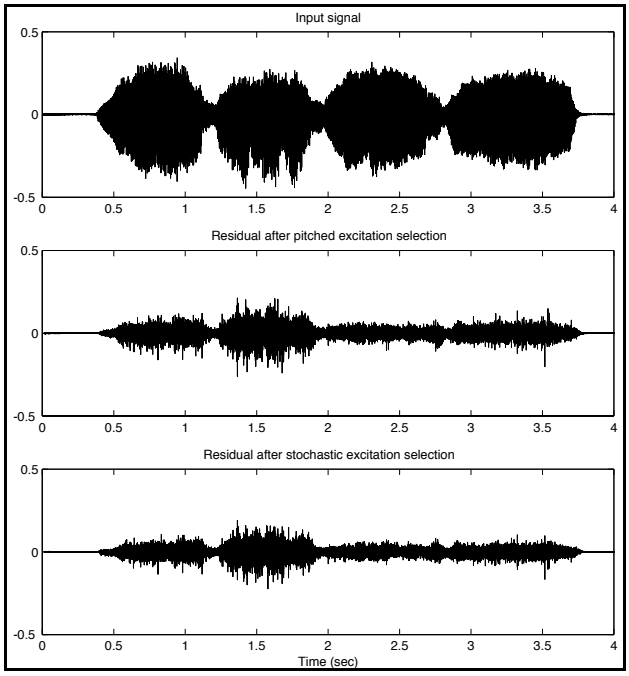

Figure 2: *Singing residuals after pitched and stochastic excitation selections.*

3.2. Codebook Improvements for the Singing Voice

The most obvious improvement to be made is to the stochastic codebook, which provides little benefit to the encoding of singing. When analyzing the residual after re-synthesis using the pitched excitation, it is apparent that there is still a large amount of harmonic material remaining. Thus, it is unlikely that an alternate stochastic codebook designed to provide noise-like excitations would do any better. Instead, a codebook derived from residual error data taken from recorded singing was used to replace the stochastic codebook. Approximately one continuous minute of sung material from an individual singer was used to create the new codebook.

The residuals after self-excitation coding of the recorded material were divided into subframes of the same length as the excitation subframes. A Principle Components Analysis (PCA) of the residual subframes was performed using singular value decomposition, and the eigenvectors corresponding to the highest 128 eigenvalues were selected for the codebook [10]. Example codebook vectors are shown below in Figure 3.

This is a much different type of codebook than the one used in the standard CELP coder. This codebook consists of basis vectors which are combined to model the individual excitations. This involves projecting each subframe vector $\boldsymbol{x}$ of the error residual onto the codebook matrix, $\boldsymbol{P}$, to find the weighting vector $\boldsymbol{w}$ for that subframe:

$$\boldsymbol{w} = \boldsymbol{P}^{\mathrm{T}} \bullet \boldsymbol{x} \tag{3}$$

$\boldsymbol{w}$ is a vector of length 128, where each value indicates the contribution of the corresponding eigenvector (codebook

vector) to $\boldsymbol{x}$. We use the eight codebook vectors with the largest contributions to model the excitation, $\boldsymbol{e}$:

$$\boldsymbol{e} = \boldsymbol{P} \bullet \boldsymbol{w}_1 \quad (4)$$

where $\boldsymbol{w}_1$ contains the eight values selected from w in their original locations. Calculation of this excitation is not as computationally intensive as the codebook search. This encoding, however, requires more bandwidth since we must now transmit eight indexes and gains per subframe instead of just one. The results of this encoding can be seen in Figure 4, where the final residual has been noticeably reduced.

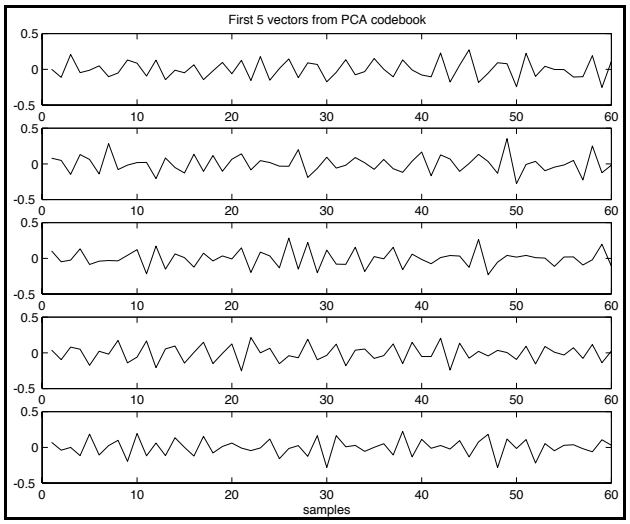

Figure 3: *Example codebook entries*

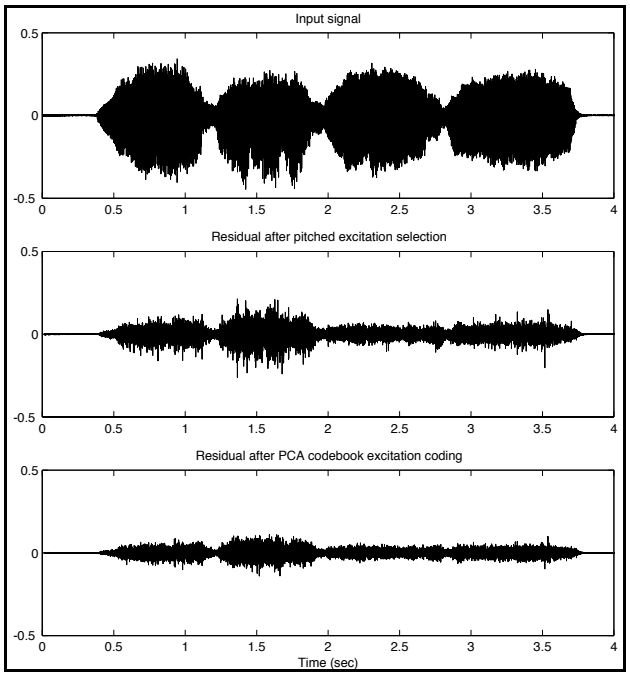

Figure 4: *Singing residuals after pitched excitation selection and PCA codebook search.*

4. FUTURE DIRECTIONS

This type of codebook results in singing encoding of noticeably better sound quality than a standard CELP implementation, though at the cost of increased bandwidth. We are currently investigating ways to reduce the extra bandwidth required. Though the research described here uses PCA to derive the codebook vector, other techniques such as Independent Components Analysis (ICA) [11] may yield better basis vectors for the codebook and will be explored further at a later date. Sound quality may also be improved by improving the pitch track, which often finds multiples of the fundamental period instead of the actual fundamental.

This research has only investigated codebook generation for individual voices. It may be possible to generalize this technique to multiple or even all voices. An obvious, though data intensive method, would be to generate code vectors from a much broader collection of data with a wide variety of voices. This will be the subject of a future investigation.

5. REFERENCES

[1] Y. E. Kim, "Structured Encoding of the Singing Voice Using Prior Knowledge of the Musical Score". *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 47-50, 1999.

[2] J. Sundberg. *The Science of the Singing Voice.* Dekalb, IL: Northern Illinois University Press, 1987.

[3] P. R. Cook. *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. Unpublished Ph.D thesis. Stanford U., 1990.

[4] B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured Audio: Creation, Transmission, and Rendering of Parametric Sound Representations". *Proceedings of the IEEE,* vol. 85, no. 5, pp. 922-940, 1998.

[5] E. D. Scheirer, "Structured audio and effects processing in the MPEG-4 multimedia standard," *Multimedia Systems*, vol. 7, no. 1, pp. 11-22, 1999.

[6] E. D. Scheirer and Y. E. Kim, "Generalized Audio Coding with MPEG-4 Structured Audio*," Proc. AES 17th Int'l Conf. on High-Quality Audio Coding*, Florence, Italy, 1999.

[7] A. Spanias, "Speech Coding: A Tutorial Review," *Proceedings of the IEEE*, vol. 82, pp. 1539-1582, 1994.

[8] J. Makhoul "Linear Prediction: A Tutorial Review," *Proceedings of the lEEE*. vol. 63, pp. 1973–1986, 1975.

[9] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[10] R. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: Wiley, 2000.

[11] A. Hyvärinen, "Survey on Independent Component Analysis," *Neural Computing Surveys*, vol. 2, pp. 94-128, 1999.