STRUCTURED ENCODING OF THE SINGING VOICE USING PRIOR KNOWLEDGE OF THE MUSICAL SCORE

Youngmoo E. Kim

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

ABSTRACT

The human voice is the most difficult musical instrument to simulate convincingly. Yet a great deal of progress has been made in voice coding, the parameterization and re-synthesis of a source signal according to an assumed voice model. Source-filter models of the human voice, particularly Linear Predictive Coding (LPC), are the basis of most low-bitrate (speech) coding techniques in use today. This paper introduces a technique for coding the singing voice using LPC and prior knowledge of the musical score to aid in the process of encoding, reducing the amount of data required to represent the voice. This approach advances the singing voice closer towards a structured audio model in which musical parameters such as pitch, duration, and phonemes are represented orthogonally to the synthesis technique and can thus be modified prior to re-synthesis.

1. INTRODUCTION

A great deal of research (primarily dealing with speech) has focused on voice coding using an analysis/re-synthesis approach. In this approach a source signal is analyzed and re-synthesized according to a source-filter model of the human voice. This is the general principle behind Linear Predictive Coding (LPC) of speech. The primary advantage of this technique over direct transmission of a recorded voice signal is the possibility of data compression, which has been the goal of most research into speech analysis/re-synthesis systems resulting in the low-bitrate speech coders used in many applications today.

This research centers around the same analysis/re-synthesis approach using LPC, but differs from traditional voice coding in several ways. First, only the singing voice, specifically Western classical singing, is considered; this allows the use of a priori musical knowledge to aid in encoding process. Second, the goal is not only data compression, but also the creation of a more flexible model for coding the singing voice that can be manipulated in a musical context. The separation of musical parameters (such as pitch, duration, and lyrics—anything defined by the musical score) from the sound generation technique is one of the core concepts of the structured audio paradigm [1]. And third, further insight will be gained into techniques that could benefit direct synthesis of the singing voice, without the requirement of a source signal.

In this paper, techniques are introduced for detecting vowel sounds in a singing-voice signal by anticipating the vowel, pitch, and duration indicated in the musical score. As the onset and release timings of vowels are detected, the LPC filter parameters during the vowel duration can be replaced by a single filter matched to the desired vowel. The resulting parameterization is more compact than LPC and still maintains comparable sound. The transitions between vowels (generally consonants) are parameterized using traditional LPC. The resulting technique is a hybrid voice coder that is both more efficient than LPC and in some ways more flexible.

2. BACKGROUND

This section presents background material for the research that follows in subsequent sections, particularly issues related to voice coding and the ways in which singing differs from speech. A definition of structured audio is presented and its relationship to voice coding is examined. The benefits of score-based analysis techniques are also discussed.

2.1. Singing voice coding

Synthesis of the human voice, speaking or singing, has proven to be an elusive goal. From an acoustic standpoint, this is primarily due to the rapid acoustic variation involved in the singing process. In order to pronounce different words, a singer must move their jaw, tongue, teeth, etc., changing the shape and thus the acoustic properties of the vocal tract. Since no other instrument exhibits the amount of physical variation of the human voice, synthesis techniques that are well suited to other musical instruments often do not apply well to singing synthesis.

The process of singing [2] starts with the breath pressure produced by the lungs. The pressure forces open the vocal folds, which are then sucked back together by the Bernoulli force. Rapid repetition of this process results in phonation, which correlates to our perception of the pitch. The shape of the vocal tract, consisting of the throat, mouth, nose, tongue, teeth, and lips, acoustically filters the source sound produced by the vocal folds. Reconfiguration of the vocal tract for different syllables creates different filters, and the output from these filters is perceived as different phonemes. Vocal sounds are characterized as voiced or unvoiced, depending on whether phonation occurs for the sound. For example, all vowels and some consonants ([m], [n], [l]) are voiced, while other consonants ([f], [s], [t]) are not. For unvoiced sounds, the source is no longer the phonation of the vocal folds, but the turbulence caused by air impeded by the vocal tract. Some consonants ([v], [z]) are mixed sounds that use both phonation and turbulence to create the overall sound.

Historically, speech and singing research have been closely linked, but in some ways, singing voice parameterization poses

an easier problem. The vast majority of sounds generated during singing are voiced whereas speech contains a much larger percentage of unvoiced sounds. Because of their periodic nature, voiced sounds are easier to analyze and generate using linear signal processing theory. In Western music, each note that is sung is fairly constant and quantized in pitch, as opposed to speech, in which pitch varies unpredictably and continuously. This simplifies the task of generating an accurate pitch track of a singing voice signal. In the most common classical singing technique, known *as bel canto*, singers are taught that vowel sounds are most efficient 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. This consistency makes it easier to determine the vowel from analysis of the signal. Most importantly for the present research, singing in the Western classical tradition is an *interpretation* of a predefined musical score. Knowledge of the score provides significant advantages in parameter extraction, in particular by showing what to look for in pitch and vowel selection.

There are several other successful approaches to singing voice coding. The homomorphic vocoder [3], which is based on cepstral analysis, was used to restore old recordings of Enrico Caruso. Another coding technique is sinusoidal analysis/ synthesis [4] in which the individual partials of the singing voice are tracked. Both of these techniques result in high-quality sound, but require more bandwidth than most LPC-based coders. Sinusoidal analysis/synthesis has also been used as the basis for a direct synthesis system for the singing voice [5]. Physically modeled synthesis of singing [6] could conceivably be used for voice coding as well, but extraction of physical parameters of the vocal tract from an acoustic signal remains problematic.

2.2. Singing in the Structured Audio Context

Vercoe *et al.* [1] coined the term *structured audio* to tie together research on the creation, transmission, and rendering of parametric sound representations, or more simply put, model-based audio. Many low-dimensional parametric instrument models exist for simulating high-quality instrument sounds that can be represented in current algorithmic structured audio implementations, such as MPEG-4 Structured Audio [7]. As of yet there is no such low-dimensional model for the singing voice. A structured singing voice model would be able to use knowledge about the music itself (the score, lyrics, etc.) for very compact representation. More importantly, the parameters in such a model could also be modifiable, creating more opportunities for interaction than with a naturally coded audio signal.

All instruments share what is known as the *encoding problem*, or the extraction of control parameters from an audio signal. The control parameters can be used in a structured representation of the music to re-synthesize the audio. LPC can also be used to alter sounds other than the human voice in a musical context [8]. One of the benefits of LPC is that parameter extraction is straightforward, the cost being that LPC is a relatively high-dimensional model. The LPC filter parameters are also difficult to relate to perceptual features, which are the kind of parameters ideally suited to low-dimensional instrument models. The coding technique presented in the next section of this paper takes steps towards defining such a relation.

Algorithmic implementations of structured audio, as in the MPEG-4 standard, can be used to emulate the behavior of other audio coding techniques, including perceptual transform coding and LPC. This technique is called *generalized audio coding* [9], and can lead to hybrid coding techniques combining aspects of traditional audio coders with the flexibility of synthesis. An example of this kind of hybrid coder, an extension of LPC, is proposed in this paper.

2.3. Score-based transcription and analysis

The use of score-based analysis in this paper is inspired by previous work by Scheirer [10] that used prior knowledge of a piano score to extract expressive performance information from recordings. Scheirer's system tracked keyboard onsets and releases based on predictions made from the score. The approach used in this paper is based upon that earlier system, with significant modifications for the acoustic properties of the singing voice versus those of the piano. In particular, no timbral model was required in the case of the piano, whereas one is needed for voice in order to identify different vowels.

3. SCORE-BASED PARAMETER EXTRACTION

The analysis model presented here takes a digitized source signal from a human singer (singing from a pre-defined score) and outputs the standard LPC parameters of pitch, gain, and filter coefficients. For simplicity, the data used for this experiment was the phrase "Alleluia, Amen", performed by a trained singer. This phrase consists of only four vowel sounds, the International Phonetic Alphabet symbols for which are [a], [e], [u], and (briefly) [i] and three liquid voiced consonants: [l], [m], and [n]. While this is a small subset of the possible phonetic choices, the techniques for vowel identification and analysis may be extended to include other vowels.
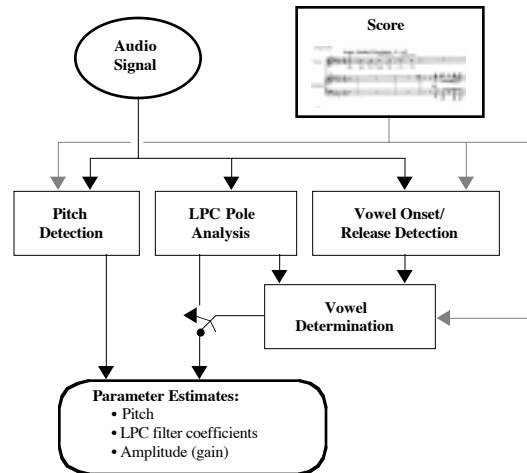

Figure 1: *Block diagram of analysis system*

3.1. Analysis blocks

The singing sources used in this experiment were digitized at a sampling rate of 11025 Hz. The system uses prior knowledge of the musical score to aid in the determination of the current pitch and vowel being sung. The parameters are estimated on a per-frame basis, where each frame is approximately 45ms in length (500 samples). Frames were overlapped by 50% and windowed using a Hanning window prior to processing.

The information presented in the score included the time signature, the tempo (beats per minute), the onset and offset of each note (in beats within a measure), and the primary vowel of each note. The score format was designed to present the same amount of information as standard musical notation.

3.1.1. Pitch Detection

Pitch extraction in this system is performed by finding the peak of an autocorrelation of the windowed signal that is targeted within a specific range defined by the score (from the current pitch period to the next pitch period). In this way, most errors are avoided, such as octave errors common to pitch detection by simple autocorrelation. This method was chosen for computational simplicity, and because autocorrelation is also used in each frame for the calculation of the LPC coefficients.

3.1.2. Linear Predictive Coder

The system implements a subset of traditional LPC techniques. As mentioned previously, *bel canto* singing consists mostly of vowel sounds, so the identification of consonant sounds is ignored in this experiment; they are simply coded using LPC. Because most singing is voiced, a voiced/unvoiced determination is not made and only entirely voiced examples were used.

The goal of linear predictive analysis is to establish 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 shown [11] to be an all-pole filter:

$$H[z] = \frac{G}{A[z]} \tag{2}$$

where the denominator is defined as follows:

$$A[z] = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{3}$$

This shows that linear predictive analysis is equivalent to a source-filter model, where the vocal tract response is modeled using a time-varying all-pole filter function of order p. The primary contributions to the filter are the resonances of the vocal tract, but also include all variations not represented by an impulse train source (radiation loss, glottal pulse shape). The derived filter will never be the true vocal tract filter, but in practice it is a reasonable approximation.

The autocorrelation method was used to establish the parameter values α_k for each analysis frame. This technique minimizes the mean squared prediction error in each frame to calculate the filter parameter values [11]. The prediction error is the difference between the source signal and the predicted signal. Thus, the squared prediction error is:

$$E_m = \sum_m \left(s_m[n] - \tilde{s}_m[n] \right)^2 \tag{4}$$

The calculated filter parameters form a polynomial, which can be factored to determine the pole locations. For low order polynomials, the pole angles generally correspond to the formant frequencies. An order of $p=8$ was used so that each pole pair would locate one of the four largest formants.

The gain parameter in each frame was calculated from the energy in the error prediction signal. This is the standard method for determining the gain [11].

3.1.3. Vowel Onset and Release Detection

Training data for vowel identification was collected by having the singer sing each of the vowels [a], [e], and [u] at seven different pitches. The LPC filter parameters were calculated and averaged for each vowel to obtain a vowel template.

The system first looks for vowel onsets by examining the energy of the input signal. An onset location prediction is calculated from the score and the indicated tempo and is used to locate a detection window. The detection window spans from halfway between the predicted onset and the previous calculated onset to the midpoint between the predicted onset and the next predicted onset. The current system is restricted to sounds without diphthongs (consecutive vowel sounds), so vowel onset will occur either at a note onset or after a consonant. The energy of a vowel is likely to be greater than the energy of a consonant because the vocal tract is open for vowels and closed for consonants. Thus, the location of the vowel onset is taken to be the local maximum derivative of the closest to the predicted onset, which accounts for both cases in which the vowel is preceded by a consonant and cases in which the vowel is preceded by silence. Calculated onsets are used to readjust the tempo estimate, which adjusts the next predicted onset.
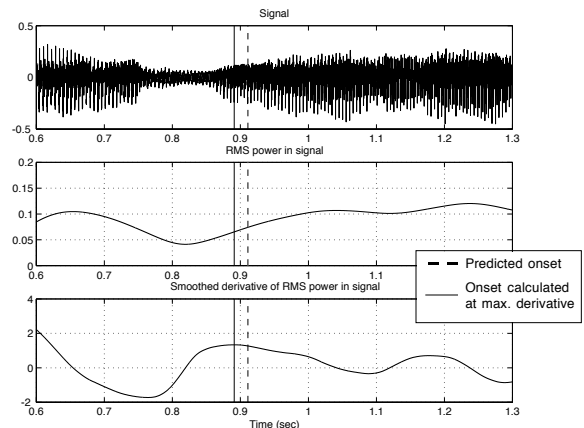

Figure 2: *Vowel onset detection of [e] in 'alleluia'.*

The vowel releases are located after all the onsets have been found. A release detection window spans from halfway between a note's predicted release and its calculated vowel onset to the calculated vowel onset of the next note or the end of the file. Again a consonant or silence follows each vowel, so the energy of the signal is used to determine the release location. The vowel release is taken to be the point at which the energy falls below 60% of the maximum energy in the note (between the times of consecutive onsets). The space between the onset and offset is the vowel duration. Space outside the vowel duration (silence and note transitions usually indicative of consonants) is encoded using standard LPC parameterization.
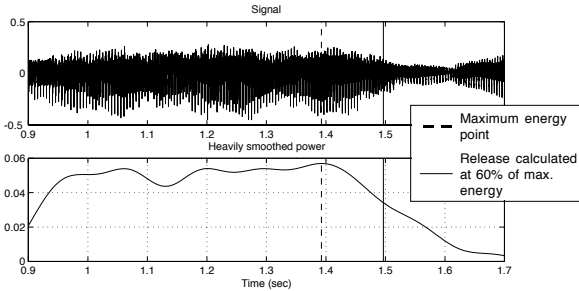
Figure 3: *Vowel release detection of [e] in 'alleluia'.*

3.1.4. *Vowel identification and compression*

Vowels within the vowel duration are identified by the locations of the formants. The formants are ordered by frequency and compared to formant locations calculated for different vowels from training data. The vowel with the smallest sum of absolute distances between ordered formants is taken to be the vowel. The frame's calculated LPC filter coefficients are then replaced with generic coefficients for the given vowel, which are also calculated from averaged training data. Since the majority of analysis frames will consist of vowels, the data required to represent a note can be greatly reduced. Of course, this is at the expense of sound quality, but the resulting re-synthesis is perceptually close to the regularly LPC coded re-synthesis.
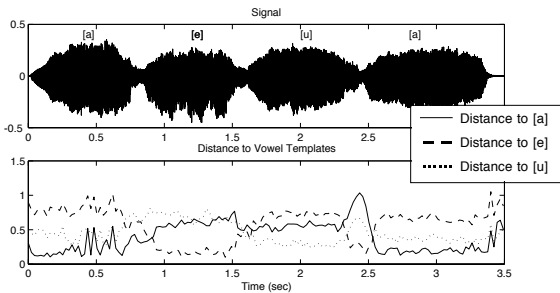

Figure 4: *Sum of formant distances to vowel templates for 'alleluia'. Smaller distance indicates a better match.*

3.2. Hybrid Coding Format

As with LPC, pitch and gain parameters are transmitted for each analysis frame. For frames in which vowels occur, no filter coefficients are needed, since they are replaced with values from the vowel templates. Thus, a codebook containing these vowel templates must also be included in the transmission. Ignoring the overhead of the codebook, this coding scheme results in a bitstream for the given sound example that is about 1/4 the size of a LPC encoded bitstream with comparable sound quality.

4. FUTURE DIRECTIONS

An obvious improvement to the current system would be to add support for the detection and synthesis of other vowels and voiced and unvoiced consonants. This would require making a voiced/unvoiced determination; there are well-documented techniques for doing that in the LPC literature, [12] for example. The increased number of choices would lead to more confusion in the detection, so a better heuristic (other than simple formant distance) for phonetic matching may be needed.

The current system could also be easily extended using different orders of LPC analysis for the vowel matching analysis and the audio analysis/re-synthesis. The current system uses a small number of poles (eight) to make formant selection, and thus vowel detection, easier. A low order LPC analysis could be used for formant detection, and a higher order could be used for the actual coding. The replacement vowel templates would also need to be recalculated at the higher order. The greater number of poles in the re-synthesis would result in better sound quality.

The techniques presented in this paper are not exclusively limited to LPC. LPC was chosen because it allows the formant frequencies to be tracked easily. However, other analysis/re-synthesis methods, such as the homomorphic vocoder [3] or sinusoidal modeling [4] could be used as the primary coding engine. Since vowel onset and release timing is calculated using the time-domain energy of the signal, it is independent of the coding technique. The difficulty would be in determining formant locations using another analysis/re-synthesis system. A peak-picking heuristic would need to be used to determine formants in either the homomorphic vocoder or sinusoidal modeling. Once the formants were determined, the vowel could be determined using the techniques outlined in the previous section. The replacement vowel templates would also need to be converted according to the synthesis technique being used.

5. REFERENCES

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

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

[3] N. J. Miller. *Filtering of Singing Voice Signal from Noise by Synthesis.* Unpub. Ph.D. thesis. Univ. of Utah, 1973.

[4] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acous., Speech, and Sig. Proc.*, vol. 34, pp. 744-754, 1986.

[5] M. W. Macon et al., "Concatenation-based MIDI-to-Singing Voice Synthesis," *Proc. 103rd AES*, Prepr. 4591, Sept., 1997.

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

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

[8] P. Lansky and K. Steiglitz, "Synthesis of Timbral Families by Warped Linear Prediction," *Computer Music Journal,* vol. 5, no. 3, pp. 45-49, 1981.

[9] 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.

[10] E. D. Scheirer, "Using musical knowledge to extract expressive performance info. from audio recordings," in D. F. Rosenthal & H. G. Olano, eds. *Computational Auditory Scene Analysis.* Mahwah, NJ: Erlbaum Assoc., 1998.

[11] J. Makhoul "Linear Prediction: A Tutorial Review," *Proc. lEEE.* vol. 63, pp. 1973–1986, 1975.

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