A SINGING VOICE SYNTHESIS SYSTEM BASED ON SINUSOIDAL MODELING

Michael W. Macon[1*] *Leslie Jensen-Link*[2†] *James Oliverio*[2] *Mark A. Clements*[1]
E. Bryan George[3‡]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250
[2]Department of Music, Georgia Institute of Technology, Atlanta, GA 30332-0456
[3]DSP Research and Development Center, Texas Instruments, Dallas, TX 75265-5474

ABSTRACT

Although sinusoidal models have been demonstrated to be capable of high-quality musical instrument synthesis [1], speech modification [2], and speech synthesis [3], little exploration of the application of these models to the synthesis of *singing voice* has been undertaken. In this paper, we propose a system framework similar to that employed in concatenation-based text-to-speech synthesizers, and describe its extension to the synthesis of singing voice. The power and flexibility of the sinusoidal model used in the waveform synthesis portion of the system [1] enables high-quality, computationally-efficient synthesis and the incorporation of musical qualities such as vibrato and spectral tilt variation. Modeling of segmental phonetic characteristics is achieved by employing a "unit selection" procedure that selects sinusoidally-modeled segments from an inventory of singing voice data collected from a human vocalist. The system, called LYRICOS, is capable of synthesizing very natural-sounding singing that maintains the characteristics and perceived identity of the analyzed vocalist.

1. BACKGROUND

Speech and singing differ significantly in terms of their production and perception by humans. In singing, for example, the intelligibility of the phonemic message is often secondary to the intonation and musical qualities of the voice. Vowels are often sustained much longer in singing than in speech, and precise, independent control of pitch and loudness over a large range is required. These requirements significantly differentiate synthesis of singing from speech synthesis.

Most previous approaches to synthesis of singing have relied on models that attempt to accurately characterize the human speech production mechanism. For example, the SPASM system developed by Cook [4] employs an articulator-based tube representation of the vocal tract and a time-domain glottal pulse input. Formant synthesizers such as the CHANT system [5] rely on direct representation and control of the resonances produced by the shape of the vocal tract. Each of these techniques relies, to a degree, on accurate modeling of the dynamic characteristics of the speech production process by an approximation to the articulatory system. Sinusoidal signal models are somewhat more general representations that are capable of high-quality modeling, modification, and synthesis of both speech and music signals [1, 2, 3]. The success of previous work in speech and music synthesis motivates the application of sinusoidal modeling to the synthesis of singing voice.

In much previous singing synthesis work, the transitions from one phonetic segment to another have been represented by stylization of control parameter contours (e.g., formant tracks) through rules or interpolation schemes. Although many characteristics of the voice can be approximated with such techniques after painstaking hand-tuning of rules, very natural-sounding synthesis has remained an elusive goal.

In the speech synthesis field, many current systems back away from specification of such formant transition rules, and instead model phonetic transitions by concatenating subword segments from an inventory of recorded speech data. These units are smoothed to diminish perceptible discontinuities at the boundaries, and time-scale and pitch modification algorithms are employed to give the speech the desired prosody [3]. With an acoustic inventory of sufficient size, this approach achieves segmental quality that approaches that of human utterances, and this motivates its exploration as a framework for singing voice synthesis.

2. SYSTEM OVERVIEW

The LYRICOS system, shown in Figure 1, uses a commercially-available MIDI-based music composition software as a user interface. The user specifies a musical score and phonetically-spelled lyrics, as well as other musically-interesting control parameters such as vibrato and vocal effort. This control information is

*MWM is currently with the Oregon Graduate Institute.
†LJ-L is currently with Momentum Data Systems, Inc.
‡This work was supported by Texas Instruments.

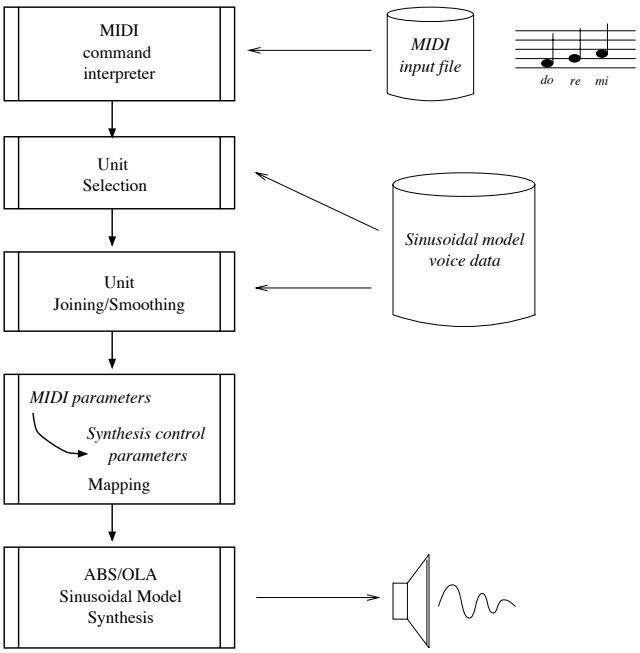

Figure 1. LYRICOS **system block diagram.**

stored in a standard MIDI file format that contains all information necessary to synthesize the vocal passage.

Based on this input MIDI file, the system selects synthesis model parameters from an inventory of voice data that has been analyzed offline by the sinusoidal model. Units are selected to represent segmental phonetic characteristics of the utterance, including coarticulation effects caused by the context of each phoneme. Algorithms described in [3] are applied to the modeled segments to remove disfluencies in the signal at the joined boundaries. The sinusoidal model parameters are then used to modify the pitch, duration, and spectral characteristics of the concatenated voice units as specified by the input musical score and MIDI control information. Finally, the output waveform is synthesized using the method described below.

3. WAVEFORM SYNTHESIS

The ABS/OLA sinusoidal model

The waveform synthesis model used is an extension of the *Analysis-by-Synthesis/Overlap-Add* (ABS/OLA) sinusoidal model [1]. In the ABS/OLA model, the input signal $s[n]$ is represented by a sum of overlapping short-time signal frames $s_k[n]$.

$$s[n] = \sigma[n] \sum_k w[n - kN_s] s_k[n] \qquad (1)$$

where N_s is the frame length, $w[n]$ is a window function, $\sigma[n]$ is a slowly time-varying gain envelope, and $s_k[n]$ represents the kth frame contribution to the synthesized signal. Each signal contribution $s_k[n]$ consists of the sum of a small number of *constant-frequency, constant-amplitude* sinusoidal components. An iterative analysis-by-synthesis procedure is performed to find the optimal parameters to represent each signal frame [1].

Synthesis is performed by an overlap-add procedure that uses the inverse fast Fourier transform to compute each contribution $s_k[n]$, rather than sets of oscillator functions. Time-scale modification of the signal is achieved by changing the synthesis frame duration, and pitch modification is performed by altering the sinusoidal components such that the fundamental frequency is modified while the speech formant structure is maintained [1].

The flexibility of this synthesis model enables the incorporation of vocal qualities such as vibrato and spectral tilt variation, adding greatly to the musical expressiveness of the synthesizer output.

Vibrato/pitch drift

The physiological mechanism of the pitch, amplitude, and timbral variation referred to as *vibrato* is somewhat in debate. However, frequency modulation of the glottal source waveform is capable of producing many of the observed effects of vibrato [6]. As the source harmonics are swept across the vocal tract resonances, timbre and amplitude modulations as well as frequency modulation take place. These modulations can be implemented quite effectively via the sinusoidal model synthesis by modulating the fundamental frequency of the components after removing the spectral envelope shape due to the vocal tract (an inherent part of the pitch modification process).

Using the graphical MIDI-based input to LYRICOS, users can draw contours that control vibrato depth over the course of the musical phrase, thus providing a mechanism for adding expressiveness to the vocal passage. A global setting of the vibrato rate is also possible. Addition of a slight nonperiodic drift of the pitch period (as suggested by [5], [7], and others) also contributes to a more human-sounding result.

Vocal effort scaling

Another important attribute of the vocal source in singing is the variation of *spectral tilt* with loudness. Crescendo of the voice is accompanied by a leveling of the usual downward tilt of the source spectrum [5]. Since the sinusoidal model is a frequency-domain representation, spectral tilt changes can be quite easily implemented by adjusting the slope of the sinusoidal amplitudes. *Breathiness*, which manifests itself as high-frequency noise in the speech spectrum, is another acoustic correlate of vocal intensity [7]. This frequency-

dependent noise energy can be generated within the ABS/OLA model framework by employing a phase modulation technique during synthesis [8].

Vocal tract length scaling

In synthesis of bass voices using a voice inventory recorded from a baritone male vocalist, it was found that the voice took on an artificial "buzzy" quality, caused by extreme lowering of the fundamental frequency. Through analysis of a simple tube model of the human vocal tract, it can be shown that the nominal formant frequencies associated with a longer vocal tract are lower than those associated with a shorter vocal tract [9]. Because of this, larger people usually have voices with a "deeper" quality; bass vocalists are typically males with vocal tracts possessing this characteristic.

To approximate the differences in vocal tract configuration between the recorded and "desired" vocalists, a frequency-scale warping of the spectral envelope (fit to the set of sinusoidal amplitudes in each frame) was performed, such that

$$\hat{H}(\omega) = H(\omega/\mu), \tag{2}$$

where $H(\omega)$ is the spectral envelope and μ is a global frequency scaling factor dependent on the average pitch modification factor. The factor μ typically lies in the range $0.75 < \mu < 1.0$. This frequency warping has the added benefit of slightly narrowing the bandwidths of the formant resonances, mitigating the buzzy character of pitch-lowered sounds. Values of $\mu > 1.0$ can be used to simulate a more child-like voice, as well. In tests of this method, it was found that this frequency warping gives the synthesized bass voice a much more rich, realistic character.

4. PHONETIC MODELING

Inventory data collection

The synthesis system presented in this paper relies on an inventory of recorded singing voice data to represent the phonetic content of the sung passage. Hence, an important step is the design of a corpus of singing voice data that adequately covers allophonic variations of phonemes in various contexts. As the number of "phonetic contexts" represented in the inventory increases, better synthesis results will be obtained, since more accurate modeling of coarticulatory effects will occur. This implies that the inventory should be made as large as possible. This goal, however, must be balanced with constraints of (a) the time and expense involved in collecting the inventory, (b) stamina of the vocalist, and (c) storage and memory constraints of the synthesis computer hardware.

To make best use of available resources, the assumption can be made that the *musical quality* of the voice is more critical than *intelligibility* of the lyrics. Thus the fidelity of sustained vowels is more important than that of consonants. Also, it can be assumed that, based on features such as place and manner of articulation and voicing, consonants can be grouped into "classes" that have somewhat similar coarticulatory effects on neighboring vowels.

Thus, a set of nonsense syllable tokens was designed with a focus on providing adequate coverage of vowels in a minimal amount of recording. All vowels V were presented within the contexts $C_L V$ and $V C_R$, where C_L and C_R are classes of consonants (e.g. voiced stops, unvoiced fricatives, etc.). The actual phonemes selected from each class were chosen sequentially such that each consonant in a class appeared a roughly equal number of times across all tokens. These $C_L V$ and $V C_R$ units were then paired arbitrarily to form $C_L V C_R$ units, which were then embedded in a "carrier" phonetic context to avoid word boundary effects.

A set of 500 inventory tokens was sung by a classically-trained male vocalist to generate the inventory data.[1] Half of these 500 units were sung at a pitch above the vocalist's nominal pitch, and half at a lower pitch. This inventory was then phonetically annotated and trimmed of silences, resulting in about ten minutes of continuous singing data used as input to the offline sinusoidal model analysis. (It should be noted that this is a rather small inventory size, in comparison to established practices in concatenative speech synthesis.)

Variable-size unit selection

Given this phonetically-annotated inventory of voice data, the task at hand during the online synthesis process is to select a set of units from this inventory to represent the input lyrics. Although it is possible to formulate unit selection as a dynamic programming problem that finds an optimal path through a lattice of all possible units based on acoustic "costs," (e.g., [10]) the approach taken here is a simpler one designed with the constraints of the inventory in mind: best-context vowel units are selected first, and consonant units are selected in a second pass to complete the unit sequence.

The method used for choosing each unit involves evaluating a "context decision tree" for each input phoneme. The terminal nodes of the tree specify *variable-size* concatenation units ranging from one to three phonemes in length. These units are each given a "context score" that orders them in terms of their agreement with the desired phonetic context, and the unit with the best context score is chosen as the unit

[1] thanks to Fay Salvaras of RKM Studios (Atlanta, GA) and Matthew Link

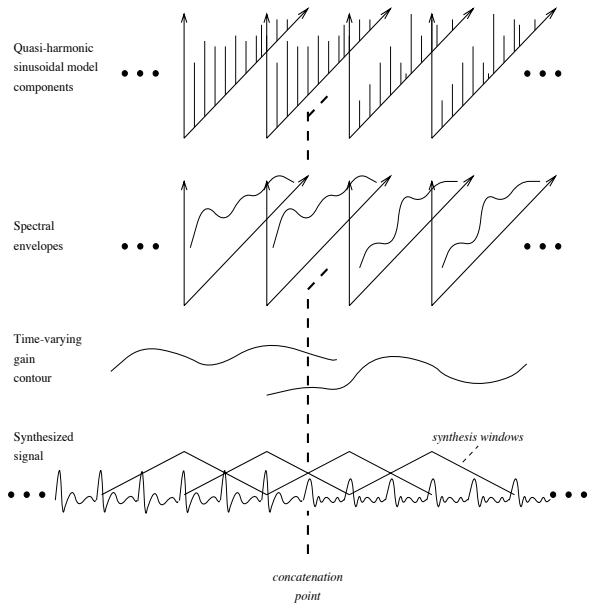

Figure 2. Concatenation of segments using the sinusoidal model.

to be concatenated. Since longer units generally result in improved speech quality at the output, the method places a priority on finding longer units that match the desired phonetic context. For example, if an exact match of a phoneme and its two neighbors is found, this triphone is used directly as a synthesis unit.

In addition to searching for phonemes in an exact phonemic context, however, the system also is capable of finding phonemes that have a context *similar*, but not identical, to the desired triphone context. For example, if a desired triphone cannot be found in the inventory, a diphone or monophone taken from an acoustically similar context is used instead. This inexact matching is incorporated into the context decision tree by looking for units that match the context in terms of phoneme *class* (as defined above). The nominal pitch of each unit is used as a secondary selection criterion when more than one "best-context" unit is available.

Once the sequence of units has been specified using the decision tree method described above, concatenation and smoothing of the units can take place, as depicted in Figure 2. Algorithms presented in [3] are used to smooth discontinuities in the spectral shape, signal energy, and phase by modifying the sinusoidal components prior to synthesis. The output waveform is then synthesized using an inverse FFT and overlap-add procedure [1].

5. SUMMARY

The system described in this paper is capable of producing a natural-sounding, musically pleasing synthetic singing voice. The LYRICOS system is novel in that it uses a "data-driven" method for modeling the phonetic information in the voice, resulting in an output that assumes the voice identity characteristics of a recorded human vocalist. It also employs a high-quality sinusoidal synthesis method capable of incorporating a wide palette of interesting musical effects into the output voice signal. A graphical input device based on an industry-standard musical instrument control language provides easy manipulation of synthesis parameters.

Demonstration via an *a capella* musical piece created with LYRICOS can be heard on the audio portion of the conference CD-ROM, and on the WWW at *http://www.ee.gatech.edu/users/macon/Sing.*

REFERENCES

[1] E. B. George and M. J. T. Smith, "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, June 1992.

[2] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, March 1992.

[3] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 361–364, May 1996.

[4] P. R. Cook, "SPASM, a real-time vocal tract physical model controller and Singer, the companion software synthesis system," *Computer Music Journal*, vol. 17, pp. 30–43, Spring 1993.

[5] G. Bennett and X. Rodet, "Synthesis of the singing voice," in *Current Directions in Computer Music Research* (M. V. Mathews and J. R. Pierce, eds.), pp. 19–44, MIT Press, 1989.

[6] R. Maher and J. Beauchamp, "An investigation of vocal vibrato for synthesis," *Applied Acoustics*, vol. 30, pp. 219–245, 1990.

[7] L. Klatt, D.H.; Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820–57, February 1990.

[8] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," accepted for publication in *IEEE Transactions on Speech and Audio Processing*, 1997.

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

[10] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 373–376, 1996.