Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models

Jordi Bonada, Òscar Celma, Àlex Loscos, Jaume Ortolà, Xavier Serra
Music Technology Group, Audiovisual Institute, Pompeu Fabra University
Barcelona, Spain
{jordi.bonada, oscar.celma, alex.loscos, jaume.ortola, xavier.serra}@iua.upf.es
http://www.iua.upf.es/mtg

Yasuo Yoshioka, Hiraku Kayama, Yuji Hisaminato, Hideki Kenmochi
Advanced System Development Center, YAMAHA Corporation
Hamamatsu, Japan
{yoshioka, kayama, hisaminato, kenmochi}@beat.yamaha.co.jp

Abstract

This paper presents an approach to the modeling of the singing voice with a particular emphasis on the naturalness of the resulting synthetic voice. The underlying analysis/synthesis technique is based on the Spectral Modeling Synthesis (SMS) and a newly developed Excitation plus Resonance (EpR) model. With this approach a complete singing voice synthesizer is developed that generates a vocal melody out of the score and the phonetic transcription of a song.

1 Introduction

The human voice is clearly the most flexible and fascinating of the musical instruments. All through the history of music, it has attracted the attention of composers and has captivated audiences. Already organ builders had the dream of imitating as faithfully as possible the sound of human voice, as we can hear in the Vox Humana stop in some organs. With the use of the new digital technologies and our current understanding of the voice, the possibility of synthesizing a natural singing voice has become more feasible.

The work presented here is a continuation of an automatic singing voice impersonator system for karaoke [Cano, Loscos, Bonada, de Boer, Serra, 2000]. That system morphed the voice attributes of a user (such as pitch, timbre, vibrato and articulations) with the ones from a prerecorded singer in real time.

The focus of our current work is to generate a performance of an artificial singer out of the musical score (melody) and the phonetic transcription (lyrics) of a song. To achieve such goal we have defined a quality evaluation criteria which takes naturalness as the fundamental consideration.

2 Introduction to Singing Voice Synthesis

Singing voice synthesis has been an active research field for almost fifty years [Cook, 1996]. Traditionally, the voice has been modeled as a linear system consisting of one or more sound sources and a set of filters which shape the spectrum of these sources. The sound source can be a periodic signal, a noisy signal, or a mixture of both, and the set of filters can be regarded as the vocal tract filters. The resulting spectrum is mainly characterized by resonant peaks called formants. Thus a vocal synthesizer has to allow the control of the resonant peaks of the spectrum and of the source parameters.

With regard to the synthesis models used in singing voice synthesis, they can be classified into two groups: Spectral models, which can be viewed as based on perceptual mechanisms, and Physical models, which can be viewed as based on production mechanisms. Any of these two models might be suitable depending on the specific requirements of the application, or they may even be combined to take advantage of both approaches.

The main benefit of using Physical models is that the parameters of the model are closely related to the ones a singer uses to control his/her own vocal system. As such, some knowledge of the real-world mechanism can be introduced in the design. The model itself can provide intuitive parameters if it is constructed with the intention that it sufficiently matches the physical system. Conversely, such a system usually has a large number of parameters. This turns the mapping of the controls of the production mechanism to the final output, and so to the listener's perceived quality, into something not trivial.

On the other hand, Spectral models are closely related to some aspects of the human perceptual mechanism. Changes

in the parameters of a spectral model can be more easily mapped to a change of sensation in the listener. Yet parameter spaces yielded by these systems are not necessarily the most natural ones for manipulation. The methods based on spectral models include Frequency Modulation, FOFs, Vocoder and sinusoidal models. Acoustic tube models are an example of physical models. Linear Predictive Coding (LPC) and formant synthesizers can be considered as spectral models and also as pseudo-physical, not strictly physical because of the source/filter decomposition they use.

Some commercial software products for singing synthesis have been released in recent years. For example, Vocalwriter [VOCALWRITER, 2000] for the English language and SmartTalk 3.0 for the Japanese language have to be mentioned [SMARTTALK, 2000]. However, the systems developed until now are far from providing enough quality to meet the practical requirements of real-world commercial applications.

The goal of a singing voice synthesis indistinguishable from a real human voice is still remote. Thus, there is a lot of room for improvement in this research area, and naturalness is one of the keywords for the work to be done. Moreover, it seems that one of the main issues behind singing voice synthesis is to offer not only quality but flexible and musically meaningful controls over the vocal sound. In that sense, we may think of applications where impossible singing voices can be synthesized or where existing voices can be enhanced.

In the next section we present the EpR voice model that we have developed and in the following ones we describe the complete voice synthesis system built around the model.

3 The EpR voice model

Our singing voice synthesizer is based on an extension of the well known source/filter approach [Childers, 1994] that we call Excitation plus Resonance (EpR). This EpR model is built on top of the sinusoidal plus residual representation obtained by the SMS analysis [Serra, 1990]. Thus the model parameters are extracted from real voice sounds that have been analyzed with SMS and in the synthesis stage the model parameters are converted to SMS parameters, from which the output sound is obtained. In this article we concentrate on the EpR model part of the method developed.

Figure 1 shows a diagram with the three types of excitation used and the corresponding filters. For the case of a voiced phonation, the filter applied to each excitation generates the appropriate spectral shape by using a frequency domain filtering function that is decomposed into two cascaded operations: an exponential decay curve plus several resonances. After filtering, the voiced residual excitation needs to be transposed to the synthesis pitch because it is a filtered SMS residual recording and has traces of the original pitch. Otherwise, in the case of an unvoiced phonation, we apply a filter that just changes the tilt curve and the gain of the STFT (Short-time Fourier Transform) of an original recording.

3.1 The EpR excitation

Voiced harmonic excitation

The inputs that control the voiced harmonic excitation generation are the desired pitch and gain envelopes. The actual excitation signal can either be generated in the time or the frequency domains. The most flexible excitation is obtained by generating a delta train in the time domain, thus allowing to achieve the best period resolution and to use some simple excitation templates. This can also be useful to generate jitter or different types of vocal disorders. This delta train can be seen as a glottal source wave previous to a convolution with the differentiated glottal pulse.

A fractional delay filter is needed to position the excitation deltas between samples, since we have to go beyond the sampling rate resolution. The filter is implemented using a windowed sinc-like function situated at each pulse location with the offset subtracted [Smith, Gosset, 1984]. Then the signal is windowed and the FFT computed. The result is a spectrum approximately flat that contains the harmonics approximately synchronized in phase.

When the harmonic excitation is generated directly is the frequency domain, which is good enough for many voiced sounds, the excitation spectrum will be perfectly flat and the phase synchronization precise.

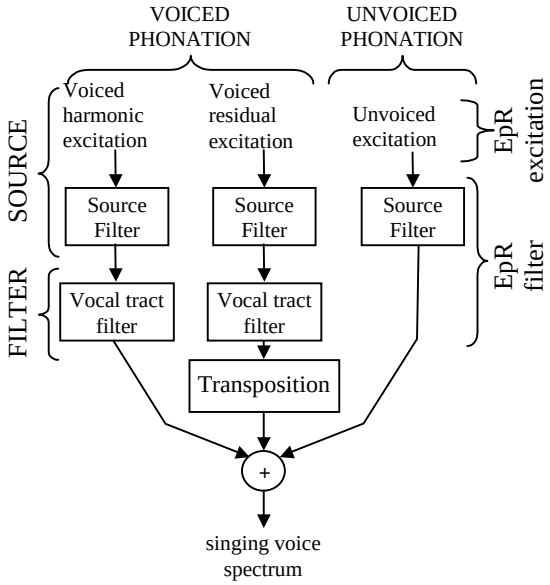

Figure 1. The EpR voice model.

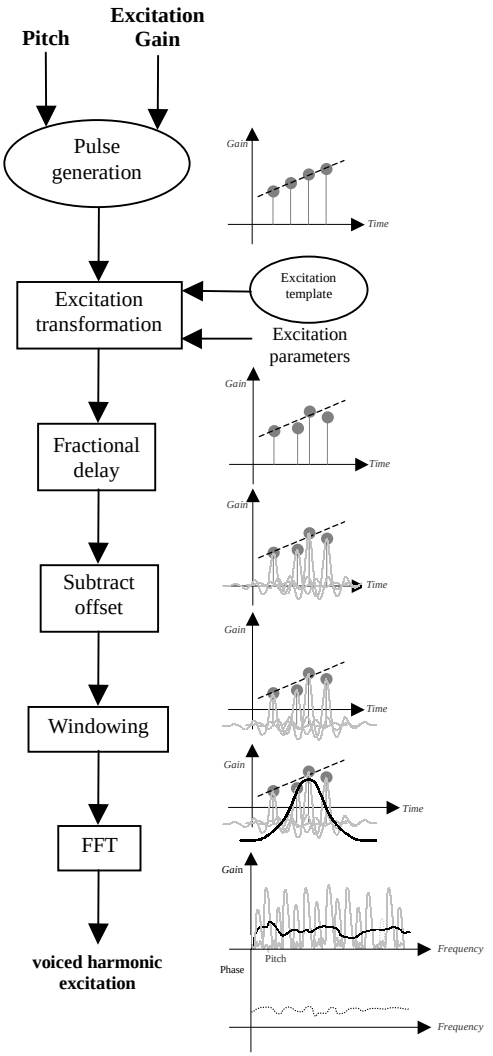

Figure 2. The EpR voiced harmonic excitation.

Voiced residual excitation
The voiced residual excitation is obtained from the residual of the SMS analysis of a long steady state vowel recorded from a real singer. The SMS residual is then inverse-filtered by its short-time average spectral shape envelope to get an approximately flat excitation magnitude spectrum.

Unvoiced excitation
The excitation in the unvoiced parts of the sounds uses directly the original recording of a singer's performance.

3.2 The EpR filter

The EpR filter is the combination of two cascaded filters. The first of them models the differentiated glottal pulse frequency response, and the second the vocal tract (resonance filter).

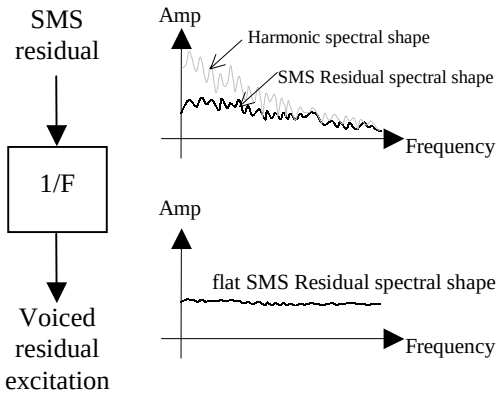

Figure 3. The voiced residual excitation.

The EpR source filter
The EpR source filter is modeled as a frequency domain curve and one source resonance applied to the input frequency domain flat excitation described in the previous section.

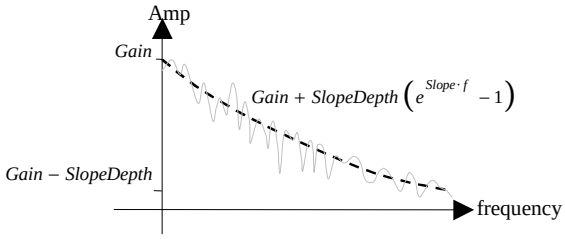

Figure 4. The EpR source curve.

This source curve is defined by a gain and an exponential decay as follows:

$$Source_{dB} = Gain_{dB} + SlopeDepth_{dB}\left(e^{Slope \cdot f} - 1\right) \quad (1)$$

This curve is obtained from an approximation to the harmonic spectral shape (*HSS*) determined by the harmonics identified in the SMS analysis

$$HSS(f) = envelope_{i=0..n}\left[f_i, 20\log(a_i)\right] \quad (2)$$

where i is the index of the harmonic, n is the number of harmonics, f_i and a_i are the frequency and amplitude of the i^{th} harmonic.

On top of the source curve, we add a second resonance in order to model the low frequency content of the spectrum below the first formant. This resonance affects the synthesis in a different way than the vocal tract resonances, as will be explained later.

The source resonance is modeled as a symmetric second order filter (based on the Klatt formant synthesizer [Klatt, 1980]) with center frequency *F*, bandwidth *Bw* and linear amplitude *Amp*. The transfer function of the resonance *R(f)* can be expressed as follows:

$$H(z) = \frac{A}{1 - Bz^{-1} - Cz^{-2}}$$

$$R(f) = Amp \frac{H\left(e^{j2\pi\left(0.5+\frac{f-F}{fs}\right)}\right)}{H\left(e^{j\pi}\right)} \quad (3)$$

where

fs = Sampling rate

$$C = -e^{-\frac{2\pi Bw}{fs}}$$

$$B = 2\cos(\pi)e^{-\frac{\pi Bw}{fs}}$$

$$A = 1 - B - C$$

The amplitude parameter (*Amp*) is relative to the source curve (a value of 1 means the resonance maximum is just over the source curve).

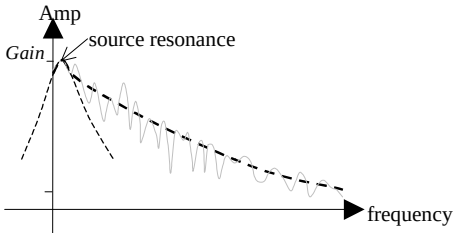

Figure 5. The EpR source resonance.

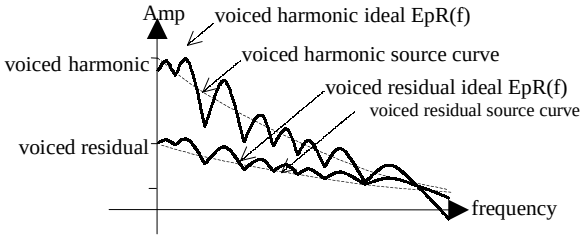

Figure 6. The EpR filter resonances.

The EpR vocal tract filter

The vocal tract is modeled by a collection of resonances plus a differential spectral shape envelope. These filter resonances are modeled in the same way as the source resonance, where the lower frequency resonances are somewhat equivalent to the vocal tract formants.

The EpR filters for voiced harmonic and residual excitations are basically the same, they just differ in the gain and slope depth parameters. This approximation has been obtained after comparing the harmonic and residual spectral shape of several SMS analysis of singer recordings. Figure 7 shows these differences.

The differential spectral shape envelope actually stores the differences (in dB) between the ideal EpR model (*iEpR*) and the real harmonic spectral shape (*HSS*) of a singer's performance. We calculate it as a 30 Hz equidistant step envelope.

$$DSS(f) = envelope_{i=0..}\left[30i, HSS_{dB}(30i) - iEpR_{dB}(30i)\right] \quad (4)$$

The EpR phase alignment

The phase alignment of the harmonics at the beginning of each period is obtained from the EpR spectral phase envelope. A time shift is applied just before the synthesis, in order to get the actual phase envelope (usually it will not match the beginning of the period). This phase alignment is then added to the voiced harmonic excitation spectrum phase envelope. The EpR spectral phase model states that each vocal tract resonance produces a linear shift of π on the flat phase envelope with a bandwidth depending on the estimated resonance bandwidth. This phase model is especially important for the intelligibility and in order to get more natural low pitched male voices.

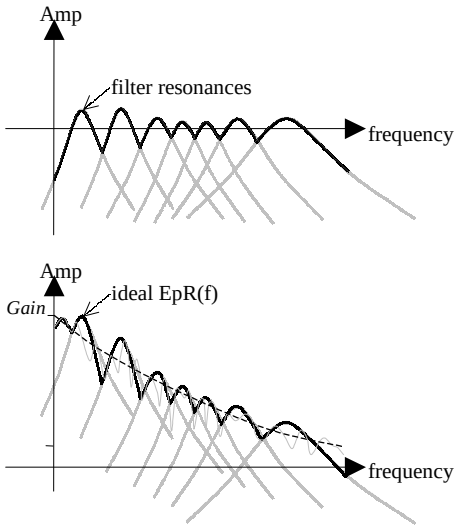

Figure 7. Differences between harmonic and residual EpR filters.

The EpR filter implementation

The EpR filters are implemented in the frequency domain. The input is the spectrum that results from the voiced harmonic excitation or from the voiced residual excitation. Both inputs are approximately flat spectrums, so we just need to add the EpR resonances, the source curve and the differential spectral shape to the amplitude spectrum. In the

case of the voiced harmonic excitation we also need to add the EpR phase alignment to the phase spectrum.

For each frequency bin we have to compute the value of the EpR filter. This implies a considerable computational cost, because we have to calculate the value of all the resonances. However, we can optimize this process by assuming that the value of the sum of all the resonances is equal to the maximum amplitude (dB) of the whole filter and excitation resonances (over the source curve) at that bin. Then we can even do better by only using the two neighboring resonances for each frequency bin. This is not a low-quality approximation of the original method because the differential spectral shape envelope, which is always kept, takes care of all the differences between the model and the real spectrum.

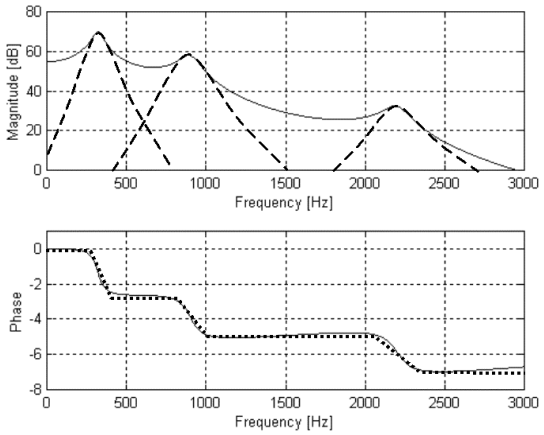

Figure 8. Phase alignment of the EpR resonances.

If we want to avoid the generation of the time domain voiced excitation, especially because of the computational cost of the fractional delay and the FFT, we can generate it in the frequency domain. From the pitch and gain input we can generate a train of deltas in frequency domain (sinusoids) that will be convolved with the transform of the synthesis window and then synthesized with the standard frame based SMS synthesis, using the IFFT and overlap-add method. However the voice quality may suffer some degradation due to the fact that the sinusoids are assumed to have constant amplitude and frequency throughout the frame duration.

The EpR filter transformation

We can transform the EpR by changing its parameters: excitation gain, slope and slope depth frequency, amplitude and bandwidth of the resonances. However, we have to take into account that the differential spectral shape is related to the resonances position. Therefore, if we change the frequency of the resonances we should stretch or compress the differential spectral shape envelope according to the resonances frequency change (using the resonances center frequency as anchor points).

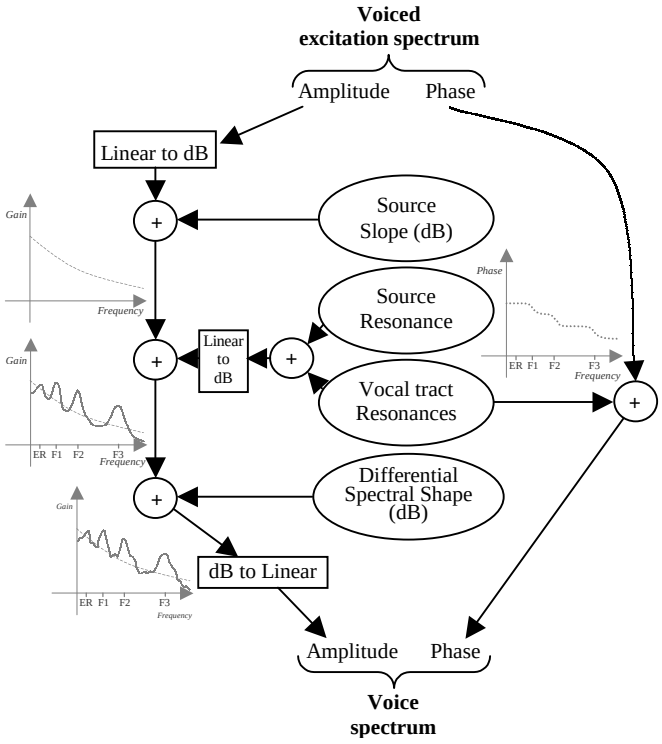

Figure 9. Frequency domain implementation of the EpR model.

4 System Overview

From the synthesis model presented in the previous section we have developed a complete singing voice synthesizer. The inputs to the system are the melody and the lyrics (in English or Japanese language), with its phonetic transcription in SAMPA format [SAMPA, 2000]. The system is able to read standard MIDI files as well as METRIX files (an ascii text file format designed as a sound synthesis control language [Amatriain, 1998]). The output of the system is a wave file with the synthesized singing voice.

The system is composed of two modules. The first one, Expressiveness module, controls the performance of the synthetic voice giving some expressiveness and naturalness to the input melody. The output of the Expressiveness module is a detailed musical score with all the needed features to characterize an expressive performance of a singer, which we call MicroScore. The second module, Synthesis module, is in charge of the actual synthesis process and uses a database containing the voice and timbre characteristics of a real singer. The analyzed voice data of

the database is the result of the SMS analysis and the EpR modelization.

5 Musical Controls

To achieve naturalness on the output voice, the system defines a set of musically meaningful controls that are either related to individual notes (note parameters) or to the whole song (general control parameters). With these parameters the system attempts to cover as many situations as possible in a singing performance. In addition to the common musical parameters (pitch, duration and dynamics), the system uses other parameters to control vocal characteristics such as attack, release, vibrato, articulation between notes, etc. The specification of all these controls has been thought of with the end-user in mind, so as to be as easy as possible to control.

The table below shows a list of the note parameters as well as the general controls. These parameters are defined in the MicroScore structure:

Note parameters	**Observations**
Pitch	Midi number (0-127) or G#3
Begin Time	of the note, in milliseconds
Duration	of the note, in milliseconds
Loudness	Normalized value [0, 1]
Lyrics	Syllable associate in a note
Dynamic Envelope	(time, normalized value)
Pitch Envelope	(time, cents)
Attack Type	{normal, sharp, soft, high}
Attack Duration	Normalized value [0, 1]
Release Type	{normal, soft, high}
Release Duration	Normalized value [0, 1]
Transition Type	{legato, staccato, marcato, portamento, glissando}
Transition Duration	Normalized value [0, 1]
Vibrato Type	---
Vibrato Depth	Envelope (time, cents)
Vibrato Rate	Envelope (time, Hz)
Opening of vowels	Envelope
Hoarseness	Envelope
Whisper	Envelope

Control parameters	**Observations**
Singer Type	Change singer of the DB
Gender Type	Change gender (male/female)
Transposition	To transpose input melody

5.1 Expressiveness Controls

In order to obtain a good synthesis, it is essential to take into account the high-level expressive controls used by real performers. In our system, this is done by the expressiveness module, which is in charge of applying certain deviations to the input score with the purpose of making it more natural and expressive.

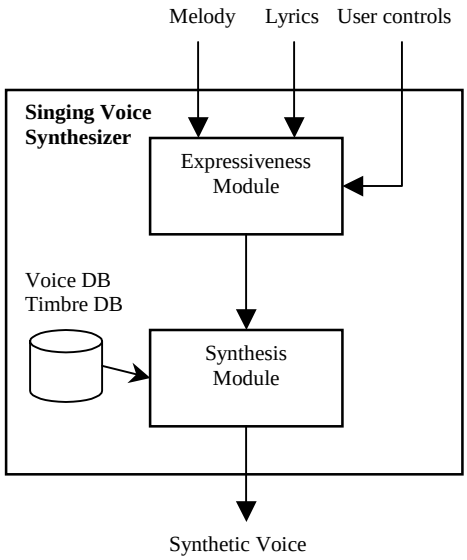

Figure 10. System overview.

Following the research made by Sundberg and Friberg on musical expressiveness [Friberg, 1991], we have implemented a very basic rule-based system for expressive control. These rules were designed as a general tool for any kind of instruments, and need adaptations and proper tuning to be suitable for the singing voice [Berndtsson, 1996]. So far, only a few rules have been successfully tested in our system. We still need more experimentation to include rules specific to voice parameters such as vibrato, breathiness, hoarseness or special kinds of attack.

An important contribution to the expressive control of sound synthesis systems has been made by Manfred Clynes [Clynes, 1987]. In view of the fact that the human ear is specially sensitive to the amplitude contour of a note, Clynes has developed a mechanism to shape individually the amplitude of each note to give more authenticity to the synthesis. In this method, called *Predictive Amplitude Shaping*, there is a global amplitude contour for every note within a musical fragment which is slightly modified for each individual note according to the pitch interval to the next note. This creates a sense of continuity and phrasing. We have implemented and tested this technique in our system, adapting the parametric controls to the human voice, with excellent results.

The pitch contour of the singing voice has to be also carefully generated in order to obtain a faithful synthesis. So we have designed a mathematical model for reproducing the smooth pitch transitions between notes. This model allows us to control the transition duration and the tuning

deviations at the end and the beginning of the notes in accordance with the musical context. In the note to note transitions, the synchronization between phonetics and musical rhythm is assured by reaching always the target pitch at the onset of the vowel of each syllable.

6 Singer Database

The creation of the database is one of the critical steps in most speech and singing voice synthesizers. There are no singing voice databases available, thus we recorded our own samples by first defining a set of musical exercises to be performed by a singer (i.e. different type of attacks, releases, note to note transitions, etc). Needless to say, the voice had to be recorded in a dry and noiseless environment to get the best possible SMS analysis.

In terms of what information is needed to be stored in the database, we can either attempt to record as many samples as possible (all possible pitches, attacks, …) or start from fewer samples and obtain the rest through transformations from the recorded/analyzed ones. This second approach gives more flexibility to the system at the possible expense of sound quality. Our approach has been this second one.

In order to choose the most useful English phonetic articulations to be included in the database, we ordered all possible articulations according to its frequency of use in actual songs. A statistical analysis was made from 76000 English songs, with close to three million words. With this information, now we know how many articulations are needed to cover a fixed percentage of all the possible articulations.

Number of articulations	Covered percentage
71	50 %
308	90 %
395	95 %
573	99 %
785	99.9 %
1129	100 %

The database is organized into two parts; timbre and voice DB.

6.1 Timbre DB

The timbre database stores the voice model (EpR) for each of the voiced phonemes. For each of these, we store several samples at different pitches and at different dynamics. When a phoneme with intermediate pitch or dynamics is required, the EpR parameters of the neighboring samples are interpolated to synthesize the phoneme at the desired pitch.

6.2 Voice DB

The voice database includes analysis data from the time varying characteristics of the voice in the form of templates. It is divided into several categories: steady states, phonetic articulations, note attacks, note-to-note articulations, note releases and vibratos. It is also possible to have different templates for the different pitches of the same type of sound.

Steady states

There are steady states stored for each of the phonemes. They model the behavior of the stationary part of a note. To do so, we store the time-varying evolution of the EpR parameters (if the phoneme is voiced) and the SMS residual along the steady state.

Phonetic articulations

All the articulations are segmented in two regions, the end of the first sound and the beginning of the next. If the regions are different in terms of voiceness, each region is analyzed with different specific parameters. Otherwise, if both regions are voiced or unvoiced the segmentation is used in synthesis to control the onset of the articulation. The EpR parameters are estimated only in the voiced part regions.

Vibratos

The vibrato templates characterize the vibrato characteristics by keeping the behavior of the voice excitation parameters. In particular, the fundamental frequency evolution, gain and source curve changes. Each time-varying function is segmented into attack, body and release. There can be different template labels according to some musical or expression classifications.

Note attacks, note transitions and note releases

These templates model the amplitude and fundamental frequency evolution at the note boundaries. They can easily be represented by a model (such as the one proposed for the amplitude by Clynes [Clynes, 1987]) which will give a smoother time evolution of the synthetic vocal sound. These templates, or models, are phoneme independent since they do not model the EpR changes. They can be organized into different expression categories.

7 Conclusions

With the purpose of demonstrating the potential of our system, two small databases with a male singer and a female singer have been created. With the female voice, we have synthesized fragments of two different songs (“We’ve only just begun”, by The Carpenters, and “Natural woman”, by Carol King). With the male voice, we have also synthesized

the same song by The Carpenters and some choruses for accompanying the female songs.

The system evaluation was made by a group of musicians that had never heard about the project. Taking the Vocalwriter synthesizer as a comparison for the English synthesis, our demo songs were considered more intelligible and more natural. In contrast, our synthesis showed a lack of timbre uniformity. As the evaluation conclusion, although the synthesis obtained was not comparable with a real singer performance, the system was judged to be of a higher quality than the available commercial systems.

From our evaluation the system still presents important drawbacks. The most important one are that unnatural artifacts appear in the synthesis, specially in the voiced consonants phonemes. Also the low register timbres of a male voice suffer from unnaturalness and sharp attacks, specially the ones belonging to plosive phonemes are smeared. We can think of some solutions to these problems by incorporating some recently proposed enhancements to the sinusoidal modeling of musical sounds [Fitz, Haken, Christensen, 2000; Verma, Meng, 2000].

The creation of a complete voice database is for now a demanding process that has to be supervised by a user. More automated ways have to be developed to facilitate and speed up the database creation process.

Certainly there is room for improvement in every step of the system. Even so, assuming, naturalness as the essential feature for evaluating the quality of a singing synthesizer, results are promising and prove the suitability of our approach.

8 References

Amatriain, X. 1998. "METRIX: A Musical Data Definition Language and Data Structure for a Spectral Modeling Based Synthesizer," *Proceedings of 98 Digital Audio Effects Workshop.*

Berndtsson, G. 1996. "The KTH Rule System for Singing Synthesis," *Computer Music Journal, 20:1, 1996.*

Cano, P.; A. Loscos; J. Bonada; M. de Boer; X. Serra. 2000. "Voice Morphing System for Impersonating in Karaoke Applications," *Proceedings of the 2000 International Computer Music Conference*. San Francisco: Computer Music Association.

Childers, D. G. 1994. "Measuring and Modeling Vocal Source-Tract Interaction," *IEEE Transactions on Biomedical Engineering 1994.*

Clynes, M. 1987. "What can a musician learn about music performance from newly discovered microstructure principles (PM and PAS)?," *Action and Perception in Rhythm and Music. Royal Swedish Academy of Music No. 55, 1987.*

Cook, P. 1996. "Singing Voice Synthesis History, Current Work, and Future Directions," *Computer Music Journal, 20:2 1996.*

Fitz, K.; L. Haken; P. Christensen. 2000. "A New Algorithm for Bandwidth Association in Bandwidth-Enhanced Additive Sound Modeling," *Proceedings of the 2000 International Computer Music Conference*.

Friberg, A. 1991. "Generative Rules for Music Performance: A Formal Description of a Rule System," *Computer Music Journal 15:2, 1991*.

Klatt, D. H. 1980. "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America, 971-995, 1980.*

SAMPA, 2000. Speech Assessment Methods Phonetic Alphabet machine-readable phonetic alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm

Serra, X.; J. Smith. 1990. "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition,'' *Computer Music Journal* 14(4):12-24.

Smith, J.O.; P. Gosset. 1984. "A flexible sampling-rate conversion method," *Proceedings of the ICASSP, San Diego, New York, vol. 2, pp 19.4.1-19.4.2. IEEE Press 1984.*

SMARTTALK, 2000. SmartTalk 3.0. Japanese Speech-Synthesis Engine with Singing Capability. http://www.oki.co.jp/OKI/Cng/Softnew/English/sm.htm

Verma, T. S.; T. H. Y. Meng. 2000. "Extending Spectral Modeling Synthesis With Transient Modeling Synthesis" *Computer Music Journal* 24:2, pp.47-59.

VOCALWRITER, 2000. Vocalwriter. Music and Vocal synthesis. http://www.kaelabs.com/