Audio Engineering Society

Convention Paper

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

Improvements to a Sample-Concatenation Based Singing Voice Synthesizer

Jordi Bonada[1], Merlijn Blaauw[1], and Alex Loscos[1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona 08003, Spain
jordi.bonada@iua.upf.edu, merlijn.blaauw@iua.upf.edu, alex.loscos@iua.upf.edu

ABSTRACT

This paper describes recent improvements to our singing voice synthesizer based on concatenation and transformation of audio samples using spectral models. Improvements include firstly robust automation of previous singer database creation process, a lengthy and tedious task which involved recording scripts generation, studio sessions, audio editing, spectral analysis, and phonetic based segmentation; and secondly synthesis technique enhancement, improving the quality of sample transformations and concatenations, and discriminating between phonetic intonation and musical articulation.

1. INTRODUCTION

Today there are around half a dozen well-known existing singing synthesizers, some of which come from academia while others are commercial products. Approaches taken to solve the problem of singing synthesis and their intended use vary greatly. Some synthesizers are for instance based on physical models of the human speech production system; others use spectral models of singing signals. Some synthesizers are specialized for operatic voices; others may be intended for pop music. In previous articles [1][2][3] we presented our synthesizer, based on real singer recording samples transformation and concatenation. One of the main features in which we put efforts ever since the beginning of the project in order to make our synthesizer unique is that synthesis should preserve the personality of the recorder singer.

This paper addresses mainly recent improvements on artificial singer’s naturalness improvement and character preservation within our synthesizer.

2. OUR SINGING SYNTHESIZER

Our synthesizer generates an artificial singing performance out of the musical score and the phonetic transcription of a song. The system stands on frame-based frequency techniques and aims to mimic human singing, for which the synthesis engine is supplied with data from a singer that has been previously recorded, analyzed and stored in a database, which stores voice characteristics (phonetics) and low-level expressivity (attacks, releases, note transitions and vibratos).

The synthesis is created by the concatenation of a set of elemental database samples (phonetic articulations and stationeries) that have been transposed and time-scaled. The concatenation of these transformed samples is performed by spreading out the spectral shape and phase discontinuities of the boundaries along a set of transition frames that surround the joint frames.

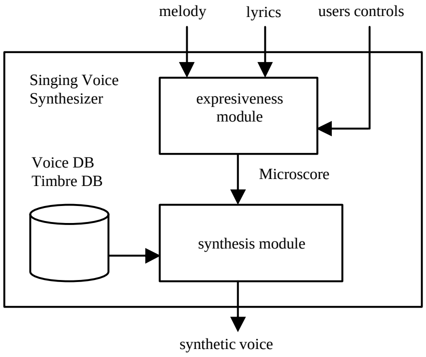

Figure 1 Block diagram of the singing synthesizer

The system can be characterized with two modules: the expressiveness module and the synthesis module. The expressiveness module gets lyrics and melody score and generates a detailed musical score that characterizes the expressivity of the virtual singer performance through an ordered list of temporal events, each of which describes the local expression of the performance through control parameters. We call this score Microscore. The Synthesis module is in charge of the synthesis process itself. The module reads the Microscore information and synthesizes the virtual performance by taking the corresponding samples from the database, transforming them according to the inputted score and concatenating them.

3. SYNTHESIS IMPROVEMENTS

Regarding the synthesis engine, three main improvements are presented here: new Microscore creation, new sample transformation technique, phonetic intonation preservation.

3.1. Improved Microscore creation

The Microscore is the internal score built from the given user input such as notes, lyrics and control curves. The score consists of a sequence of samples and corresponding sample transformations that are required to achieve the target synthesis output. The goal when creating the score is to minimize the required sample transformations, which can result in artifacts in the synthesis output. Ideally the total of all required transformations of the entire score are minimized, but in real-life implementations the time-interval over which this optimization is performed must be limited. In our current implementation it is fixed to the duration of two succeeding notes. While bigger optimization intervals get closer to the ideal solution, they also result in greater computational costs while only offering diminishing improvements. Furthermore in real-time settings, the interval is limited by the available time between user input and synthesis output (system's latency).

The improved internal score generation is based on minimizing a sample-target mismatch error function rather than the discrete decision tree of previous approaches. This avoids the decision order dependencies of the old system and is much more flexible and extensible.

In order to be able to minimize the error function, the error for all of the possible sample sequences has to be computed. Fortunately by eliminating certain cases, the complexity of this computation can be reduced to an acceptable number. The error function itself is determined by the transformations required to match pitch, dynamics and duration of the sequence of database samples to synthesis target. Additionally the continuity of sample sequence with respect to pitch and dynamics also affects the total error to avoid samples from very different contexts getting concatenated.

$$\begin{aligned} Err_{total} = Err_{pitch} + Err_{dynamics} + Err_{fitting} + \\ Err_{pitch\_continuity} + Err_{dynamics\_continuity} \end{aligned} \qquad (1\,)$$

Another major improvement to the internal score creation, besides the overall adaptation toward a multi-lingual system, is the use of additional markers in the articulation samples which determine which parts of the sample is stable and which parts are transitional. For instance an articulation sample may include a sustained vowel part or a small silence before a plosive. Using this information the system can determine which parts of the sample are essential and which can be cut without loss of intelligibility. In many cases this allows to avoid time-compressing samples when synthesizing fast singing without having dedicated rapidly sung samples in the database.

3.2. Sample Transformations

In our previous work voice transformation was accomplished by means of different spectral processing techniques based on sinusoidal modeling [1] and phase-locked vocoder [2], with the addition of a voice model we call EpR [1], which is a parameterized source-filter spectral model of the singing voice consisting of resonances, excitation curve and a residual spectral envelope. Higher quality results can be obtained by combining EpR with a technique based on non-linearly scaling the spectrum.

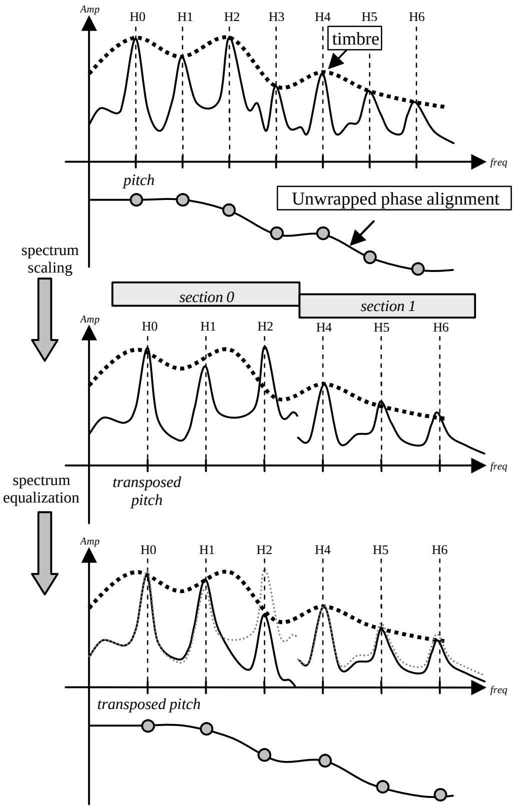

Figure 2 Hand drawn representations of spectrum scaling and equalization that take place in transposition. Unwrapped phase alignment is drawn at the voice pulse onset

The scaling procedure is actually performed in a way that the local polar spectrum shape of each harmonic is approximately shifted to new amplitude, frequency and phase coordinate. The voice signal during voiced utterances is mainly composed of harmonics and noise, and both components should be transformed in a natural way. Hence, since harmonics and surrounding noise are shifted together in frequency, spectral regions should be preferably shifted only within a small frequency range. This can be achieved by defining a set of frequency bands and a mapping between input and output spectrums which minimizes the frequency shifting of each band. In the simplest case each band would correspond to a single harmonic, and the source for each target harmonic would be the closest harmonic in the original signal. In figure 2 we can see an example of transposition upwards and frequency bands with a width equal to three harmonics.

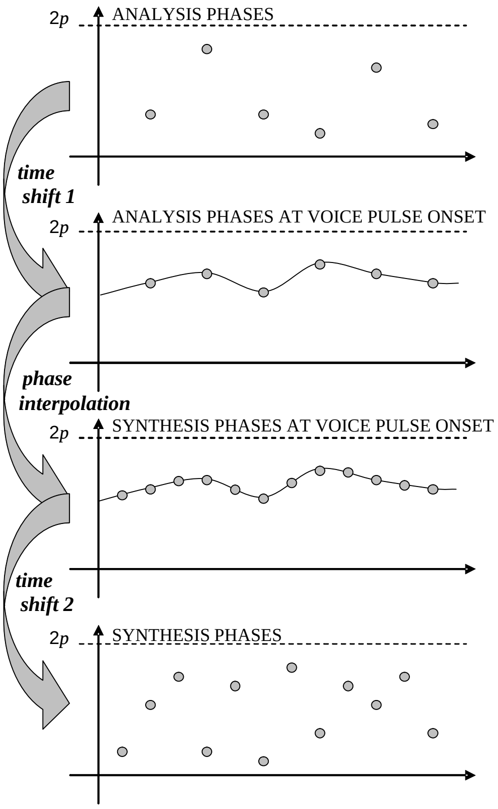

Figure 3 Hand drawn representation of the computation of synthesis phases by means of the MFPA algorithm

In all previous modeling approaches phasiness artifacts (a lack of presence, a slight reverberant quality) appear often producing unnatural results. The reason resides in the loose of the phase synchronization (or phase-coherence) between the various harmonics inherent to the voice utterances. Several algorithms have been proposed regarding this issue [4][5], most based on the idea of defining pitch-synchronous input and output onset times and reproducing at the output onset times the phase relationship existing in the original signal at the input onset times. However, the results are not good enough because the onset times are not synchronized to the voice pulse onsets, but assigned to an arbitrary position within the pulse period.

We proposed in [6] a method to estimate the voice pulse onsets out of the harmonic phases based on the property that when the analysis window is properly centered, the unwrapped phase envelope is nearly flat with shifts under each formant, thus being close to a Maximally Flat Phase Alignment (MFPA) condition. Since we process at a constant frame rate, for each analysis frame we can estimate the time distance between the center of the analysis window and the closest MFPA position (i.e. voice pulse onset) and rotate the harmonic phases to that time. Next we unwrap the harmonic phase envelope and compute the synthesis harmonic phases at the pulse onset using interpolation. Finally, we apply a linear phase shift which rotates the harmonic phases at the actual synthesis position. The whole procedure is shown in figure 3. By means of this technique, phasiness can be greatly reduced to be almost inaudible and transformations sound more natural.

3.3. Phonetic Intonation versus musical articulation

In our recording scripts we set tempo, pitch and loudness to be constant along each sentence. The main reason is that we want to capture loudness and pitch variations inherent to phonetic articulations (see figure 4), and make them independent of the ones related to musical performance. This is especially important to get intelligible and natural sounding outputs.

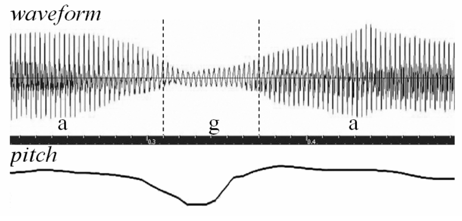

Figure 4 Example of typical amplitude and pitch envelopes in phonetic transitions

We have modified the synthesis engine in such a way that pitch and loudness envelopes from database samples are added on top of the ones computed out of the input notes and expression. With this, pitch and dynamic variations due to phonetics intonation is well preserved and the naturalness of the synthesis is significantly improved.

4. DATABASE CREATION IMPROVEMENTS

The singer database holds two different types of information: the phonetic information and the expression information. The phonetic database stores the singer's timbres in all possible contexts and the expression database stores the singer's expression in different musical contexts.

4.1. Recording Sentences

As already suggested in section 3.2, improved recording scripts are compound of sentences instead of words. Recoding scripts based on sentences ease the comprehension of the required performance, reduce the required recording time, and enhance the consistency of the pronunciation.

A script with phonetic transcription capabilities is in charge of choosing from a collection of digital books, the minimum set of these sentences required to cover all articulations to be recorded. As an example, for the Spanish recordings, we consider 29 possible allophones from [7], which makes a list of 841 theoretically possible allophone combinations. From this list impossible combinations are removed, impossible meaning they have never appeared in the automatic phonetic transcription of our book collection. Current scripts contain 521 allophone to allophone articulations in 115 sentences, covering around 94% of all possible appearing combinations, and summing around 99% of occurrence.

4.2. Diphoneme Level Segmentation

Segmenting recorded utterances into diphoneme samples was previously done using an available Automatic Speech Recognition (ASR) toolkit which uses HMM-based models trained with mel-frequency cepstral data [3]. While overall this approach gives fairly good results, the main problem is that segmentation still fails in a significant number of cases. In general, segmentation either fails completely, probably due to fundamental differences between the singing voice input and speech corpus the ASR toolkit is trailed with, or, more commonly, the overall segmentation is correct but the boundaries are mis-estimated locally. This latter is worsened by the correct segmentation being defined slightly differently between the ASR toolkit and the synthesizer.

To improve the results of the ASR toolkit a second, post-processing step was later introduced. First a number of low-level descriptors such as amplitude envelope, delta mel-cepstrum coefficients, zero-crossing rate, etc. are computed from the recording. Then, assuming the initial ASR segmentation is at least globally correct, the algorithm uses the low-level descriptors to improve the initial segmentation results using rules based on a priori knowledge the given phoneme types for an articulation. This post processing step significantly improves the results of the ASR segmentation, but has the disadvantage that finding a set of rules that work well in all cases is very difficult and still requires manual verification that the initial segmentation is at least globally correct.

4.3. EpR estimation from templates

In order to improve the automatic EpR estimation performed as part of the singer database creation process we have built up a table in which the expected frequency positions of the first to forth resonances is specified. Values in this template table have been computed from the observation of manually supervised complete singer databases.

PhU	F1	F2	F3	F4
@	700	1350	2750	3700
V	750	1250	2750	3750
e	750	1800	2600	3850
I	550	1875	2525	3700
i:	400	2225	2725	3600
{	775	1675	2675	3725
O:	700	1250	2650	3650
Q	800	1200	2675	3775
U	425	1350	2750	3650
u:	375	1550	2500	3450
@r	400	1450	2275	3450

Table 1 Expected center frequencies (in Hz) of EpR first four resonances for English voiced vowels.

The expected frequency allocation is defined by means of a center frequency and a frequency deviation (mean and variance). Variances take different values depending on the phonetic group and the context of the frame to be estimated in case of articulations. More precisely, resonances variances range from 50 Hz for voiced consonants, to 75 for vowels and diphthongs and 150 for voiced-unvoiced boundary frames.

4.4. On the fly database creation

As mentioned, the improved diphoneme segmentation algorithm works quite well in general, but has its own problems and is somewhat of an ad hoc solution. A possible way around these problems may be to utilize the existing correctly segmented databases to segment new databases. This would avoid the dependency on a speech trained model from the ASR used.

The problem of time-aligning two utterances of the same sentence by different speakers can be solved using the Dynamic Time Warping (DTW) algorithm. It takes one or a combination of descriptors of the signals and then finds the optimal path through a similarity matrix of the descriptors of both utterances. Allowing to use additional descriptors besides the mel-frequency cepstrum can achieve some of the benefits of the previous approach's post-processing step in a simpler manner because none of the phoneme-dependent rules are present. In particular, a combination of mel-frequency cepstrum coefficients and envelope derivative turned out to be an effective combination. The mel-frequency cepstrum gives a good overall match and the envelope derivative can help avoid discontinuous jumps in the optimal DTW path in cases where vowels and sustained or shortened compared to the model utterance. Another important issue with the DTW technique is to accurately trim silence at the beginning and, less importantly, the end of the utterance to remove leading and trailing silence as begin and end points of the DTW path are always fixed.

Initial listening tests to evaluate the algorithm were positive. Quantitative evaluation proved somewhat difficult because the second, reference database was only partially corrected manually. Furthermore defining a meaningful rule to measure if segmentation was successful is problematic because this is dependent on phoneme types to a degree. The inherent disadvantage of this system is of course that it requires at least one correctly segmented database for each supported language. As a consequence, the recording scripts can not be easily changed. However, creating the initial segmentation model for a specific language can still be partially automated using the previous ASR-based technique.

Another problem that arose from the way databases were created was that issues with the recordings such as mispronounced phonemes, level mis-matches, pitch problems, etc., usually weren't found until creating or using the database, when it is usually too late to fix them.

To reduce these kind of problems, the DTW-based database creation tool was implemented as a real-time VST plug-in. This allows it to be easily integrated with many recording environments. Besides improving the database creation work-flow overall, this on-the-fly system can also help flag problems as they happen and will hopefully increase database consistency which ultimately determines the synthesizer output's quality. Firstly recordings are segmented in utterances automatically, then checked if their duration is approximately equal to those of the models and finally the DTW is applied. The total error of the DTW path finding algorithm can indicate problems with the match such as severe mis-pronunciations. This system also allows levels of stationaries and vowels in articulations to be matched more closely.

5. CONCLUSIONS

Significant improvements have been achieved in the quality of the synthesis with the synthesizer modifications we have presented in this article. New strategies on Microscore creation, phonetic intonation preservation, and MPFA are key issues when it comes to preserve the identity and expression of the recorded singer.

While the improvements of the database creation and the automation thereof does not result in a better synthesis quality itself, it does allow elaborate databases to be created more easily. Sampling a singer at more different contexts such as different pitches, dynamics, and types of expression does help greatly capturing the singers characteristics and ultimately improving the quality and naturalness of the synthesis output.

6. ACKNOWLEDGEMENTS

This work was partially supported by YAMAHA´s Center for Advanced Sound Technologies, Hamamatsu, Japan.

7. REFERENCES

[1] Bonada, J., Loscos, A., Cano, P., Serra, X., Kenmochi, H., “Spectral Approach to the Modeling of the Singing Voice”, Proceedings of the 111th AES Convention, New York, USA, Sept 2001.

[2] Bonada, J., Loscos, A., Kenmochi, H., “Sample-based Singing Voice Synthesizer by Spectral Concatenation”, Proceedings of the Stockholm Music Acoustics Conference, Stockholm, Sweden, 2003.

[3] Bonada, J., Loscos, A., Mayor, O., Kenmochi, H., “Sample-based Singing Voice Synthesizer using Spectral Models and Source-Filter Decomposition”, Proceedings of 3rd Intl. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Firenze, Italy, 2003.

[4] Laroche, J., “Frequency-Domain Techniques for High-Quality Voice Modification”, Proc. of the 6th Int. Conference on Digital Audio Effects, London, UK, September, 2003.

[5] DiFederico, R., “Waveform Preserving Time Stretching and Pitch Shifting for Sinusoidal Models of Sound”, Proc. of the 1st Int. Conference on Digital Audio Effects, Barcelona, Spain, November, 1998.

[6] Bonada, J., “High Quality Voice Transformations based on Modeling Radiated Voice Pulses in Frequency Domain”, Proceedings of the 7th Conference on Digital Audio Effects, Naples, Italy, Oct 2004.

[7] Llisterri J., J.B. Mariño, ‘Spanish adaptation of SAMPA and automatic phonetic transcription’, Report SAM-A/UPC/001/V1, 1993.