Sample-based singing voice synthesizer by spectral concatenation

Jordi Bonada, Alex Loscos

Music Technology Group, Audiovisual Institute, Pompeu Fabra University
Barcelona, Spain
jordi.bonada@iua.upf.es, alex.loscos@iua.upf.es
http://www.iua.upf.es

ABSTRACT

The singing synthesis system we present generates a performance of an artificial singer out of the musical score and the phonetic transcription of a song using a frame-based frequency domain technique. This performance mimics the real singing of a singer that has been previously recorded, analyzed and stored in a database. To synthesize such performance the systems concatenates a set of elemental synthesis units. These units are obtained by transposing and time-scaling the database samples. 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. The expression of the singing is applied through a Voice Model built up on top of a Spectral Peak Processing (SPP) technique.

1. INTRODUCTION

The voice is considered to be the most flexible, rich and powerful of all instruments and it has always been an inexorable source of inspiration for musicians, sound designers and audio programmers. Music computer people have been interested in vocal sound manipulation and synthesis ever since the appearing of the very first techniques that allowed such sort of processing. Nowadays this interest has spread to most of the music communities reaching a point in which it is hard to find a popular musical production without any vocal alienation.

The synthesis of voice has been approached form many different directions and though it is a slack categorization, we can split synthesis models into two groups: spectral models, which are based on the perception of the sound, and physical models, which are based on the production of the sound.

Both spectral and physical models have their own benefits and drawbacks [1]. Physical models such as acoustic tube models use the control parameters humans use to control their own vocal system and can generate time-varying behaviors by the model itself. On the other hand some parameters might not have intuitive significance and might interact in non-obvious ways. Spectral models such as those based on phase vocoders or sinusoidal tracks have precise analysis / synthesis methods and work close to the human perception but their parameterization is not that suitable for study, manipulation, or composition since they don't match any auditory system structure.

Models such as the one used in formant synthesizers are considered to be pseudo-physical models because even though these are mainly spectral models they make use of the source / filter decomposition. The singing synthesizer we present would be part of this model group.

2. SYSTEM DESCRIPTION

The system is composed of two modules: the expression module, which is in charge of giving expressiveness and naturalness to the input melody, and the synthesis module, which is in charge of the synthesis process itself.

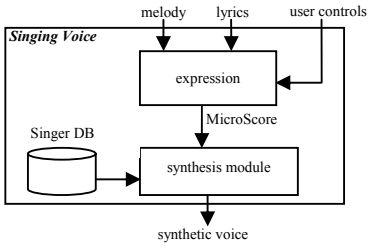

Figure 1: *General system diagram*

The indispensable inputs of the expression module are the lyrics and the musical score of the voice melody, that is to say, the phonetics and the fundamental frequency, dynamics and quantized duration values of each note. Out of these, the synthesizer can generate a virtual performance with default expression parameters. However, user control parameters can also be inputted to modify the expression through specifications on different types of note attacks, note transitions, vibratos, etcetera. The output of the expression module is 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 the control parameters. This score is called Microscore.

The inputs of the synthesis module are the Microscore and the singer database. The synthesis 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 [2].

3. VOICE AND SPECTRAL MODELING

In this section we will introduce both EpR Voice Model and Spectral Peak Processing (SPP) technique on which the system is based.

3.1. Voice Model (EpR)

Our Voice Model is based on an extension of the well known source/filter approach [3] we call EpR (Excitation plus Resonances). The EpR filter can be decomposed in two cascade filters. The first of them models the differentiated glottal pulse frequency response and the second the vocal tract (resonance filter).

The EpR Source filter

The EpR source is modeled as a frequency domain curve and one source resonance. The curve is defined by a gain and an exponential decay as follows:

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

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

$$HSS(f) = envelope_{i=0..n-1}\left[f_i, 20\log(a_i)\right]$$

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 curve, we add a resonance in order to model the low frequency content of the spectrum below the first formant. This source resonance is modeled as a symmetric second order filter (based on the Klatt formant synthesizer [4]) with center frequency *F*, bandwidth *Bw* and linear amplitude *Amp*, which is relative to the source curve.

The EpR vocal tract filter

The vocal tract is modeled by a vector of resonances plus a differential spectral shape envelope. It can be understood as an approximation to the vocal tract filter. These filter resonances are modeled just like the source resonance, where the lower frequency resonances are somewhat equivalent to the vocal tract formants.

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

$$DSS(f) = envelope_{i=0..}\left[30i, HSS_{dB}(30i) - iEpR_{dB}(30i)\right] \qquad (1)$$

Thus, the original singer's spectrum can be obtained if no transformations are applied to the EpR model.

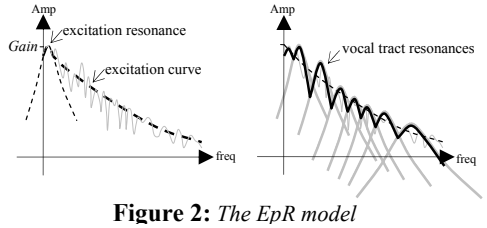

Figure 2: *The EpR model*

3.2. Spectral model (SPP)

Spectral Peak Processing (SPP) is used as the sample transformation tool. It performs a frame based spectral analysis of audio, giving as output the STFT, the harmonic peaks and the pitch. This technique is based on the phase-locked vocoder [5], but spectral peaks are calculated by parabolic interpolation [6] and a pitch analysis is performed. SPP considers the spectrum as a set of regions, each of which belongs to one harmonic peak and its surroundings. The goal of such technique is to preserve the local convolution of the analysis window after transposition and equalization transformations. A basic diagram can be seen in Figure 3.

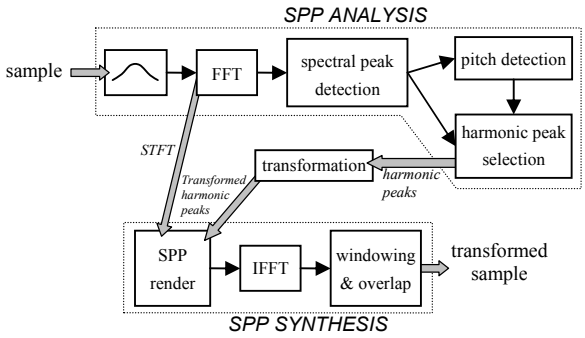

Figure 3: *SPP sample transformation*

4. SINGER DATABASE

The singer databases are created from dry and noiseless recordings of real singers. These databases hold both phonetic and expression information about the singer.

The phonetic database represents and stores the time varying phonetic characteristics of the singer so it stores the singer timbres in most representative phonetic contexts. It contains steady-states and articulations at different pitches.

The expression database characterizes the low level expression of the singer so it tries to retain the singer expressivity in different musical contexts. This database includes vibratos and note templates, which model attacks, releases and note transitions at different pitches. They are considered phoneme independent and organized into a musical meaningful classification [7].

5. TRANSFORMING SAMPLES

Basically, three kinds of transformation are applied to the samples: transposition, equalization and time-scaling. Time-scaling issues are described in detail in [8].

5.1. Transposition

Transposition means to multiply the harmonic's frequencies by a constant value. In terms of SPP, this operation can be done by shifting SPP regions in frequency. The linear frequency displacement for all the bins of each region will be the same. In most cases, this frequency displacement will be a non integer value. Thus we interpolate the spectrum with a 3rd order spline. This might not be the best method but it is a good compromise between quality and computational cost.

In Figure 4, we can see an example of transposition to a higher pitch. In this case, the SPP regions will be separated in the

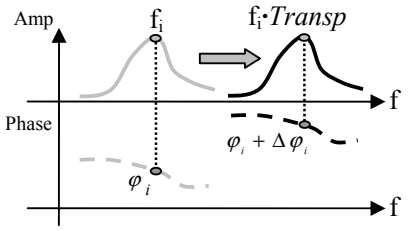

Figure 4: *Phase shift in SPP transposition*

resulting spectrum by gaps. Whenever transposing to a lower pitch, the SPP regions overlap in the resulting spectrum. This overlapping is computed by adding the complex values at each spectral bin.

Phase correction

When a SPP region is shifted in frequency in a frame by frame process, the phase needs to be corrected in order to continue the harmonics. For the i[th] harmonic, assuming linear frequency, the ideal phase increment between two consecutive frames is

$$\Delta\varphi_i = 2\pi f_i \Delta t$$

where the time increment between frames is Δt . If we transpose a sample by a factor *transp*, then the ideal phase increment should be

$$\Delta\varphi_i = 2\pi f_i \cdot transp \cdot \Delta t$$

Therefore, the amount of phase that should be added to the spectrum phase in order to continue the i[th] harmonic is

$$\Delta\varphi_i = 2\pi f_i \cdot transp \cdot \Delta t - 2\pi f_i \Delta t = 2\pi f_i \left(transp - 1\right)\Delta t$$

This phase increment is added to all the bins that belong to the i[th] region (see Figure 4). This way we can preserve the phase consistency across bins, i.e. the vertical phase coherence [5].

The phase increment applied to each harmonic is added to an accumulated phase frame by frame. This accumulated phase is the phase we finally add to each SPP region. However, this implementation results in the loss of the harmonic phase alignment after several frames because the frequencies do not follow a perfect harmonic scale. This loss of phase alignment produces some phasiness and loss of presence in the synthesized voice, especially for low pitches. To solve such problem, we consider the voice as perfectly harmonic. In that case, the equation of the phase increment will be

$$\Delta\varphi_i = 2\pi \cdot pitch \cdot \left(i+1\right)\left(transp - 1\right)\Delta t \qquad (2)$$

where *i* is the index of the harmonic (*0* for the fundamental) and *pitch* is the estimated fundamental frequency (different from f_0).

5.2. Equalization

Equalization means timbre change. On one hand the SPP analysis outputs a spectrum with the harmonic peaks and the SPP regions. On the other hand we have an envelope that defines the desired timbre. The SPP equalization is done by calculating for each region the amplitude amount needed to match the timbre envelope, and adding it to all the bins that belong to the region. Therefore, the spectrum amplitude of a region will be just shifted up or down and the phase will not be changed.

6. CONCATENATING SAMPLES

When connecting transformed samples there will appear spectral shape and phase discontinuities. In order to minimize them, phase and amplitude corrections can be spread out along a set of transition frames that surround the joint frame.

6.1. Phase concatenation

In order to avoid phase discontinuities at the segment boundaries, we have come out with a phase continuity condition that takes

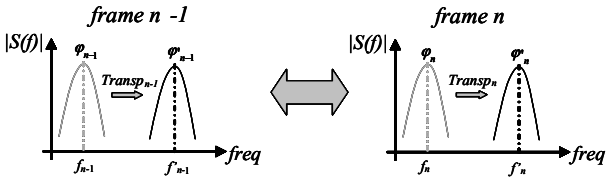

Figure 5: *Concatenation boundary*

care of the boundary phases and the possibly different transposition factors applied to each segment. In Figure 5 we can see the joint frames, where *frame n-1* is the last frame of the left segment and *frame n* is the first frame of the right segment. f_{n-1} and f_n refer to the frequencies of the i^{th} harmonic.

The basic condition for phase continuity comes from the assumption that the frequency of each harmonic varies linearly between these two consecutive frames. In that case the phase relation between both frames should be

$$\varphi_n^i = \varphi_{n-1}^i + 2\pi \frac{f_{n-1}^i + f_n^i}{2}\Delta t$$

where φ_n^i and φ_{n-1}^i are the phases of the i^{th} harmonic at the right and left frame respectively. Thus, the desired phase for the left frame ϕ_{n-1}^i should be

$$\phi_{n-1}^i = \varphi_n^i - 2\pi \frac{f_{n-1}^i + f_{n-1}^i}{2}\Delta t$$

But in fact we are not just concatenating two segments, but also transposing them with different transposition factors ($transp_{n-1}$ and $transp_n$). We should distinct then between the original frequency of each segment (f_{n-1} and f_n) and the transposed ones (f''_{n-1} and f''_n), where $f''_{n-1}=f_{n-1}\cdot transp_{n-1}$ and $f''_n=f_n\cdot transp_n$. The basic condition should be applied to the transposed frequencies and phases. This can be expressed as

$$\phi_{n-1}^i = \varphi_n^i - 2\pi \cdot \frac{f_{n-1}^i \cdot transp_{n-1} + f_n^i \cdot transp_n}{2}\cdot \Delta t + 2\pi\left(i+1\right)\Delta t \cdot pitch_n \cdot \left(transp_n - 1\right)$$

by using the transposition phase correction equation (2) and taking into account that the phase correction due to transposition is accumulated frame by frame. This correction is applied either to the left or to the right segment around the boundary, and spread along several frames in order to get a smooth transition. We can rewrite the previous equation as

$$\phi_{n-1}^i = \varphi_n^i - \Delta\varphi_c \qquad (3)$$

where $\Delta\varphi_c$ is the phase correction that guarantees the phase continuation of the i^{th} harmonic. In Figure 6 we can see an example where this phase correction is spread along 5 frames on the left part of the boundary.

Since the impulsive periods of left ant right segments are not aligned, we will often have big phase correction values. Therefore it's better if we calculate how much we should move the right segment (Δt_{sync}) in order to align the beginning of both periods, so that the phase correction to be applied is minimized. We could approximate this time shifting by assuming that the beginning of

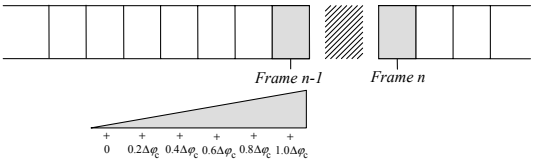

Figure 6: *Phase spreading in concatenation*

the periods will be aligned if the phase of the fundamental is continued. This can be expressed as

$$\Delta t_{sync} = \frac{-\Delta\varphi_c}{2\pi\, pitch_n \cdot transp_n} \tag{4}$$

where $\Delta\varphi_c$ is calculated as in equation (3) for the particular case of i=0 (the fundamental). Finally, if we combine equations (3) and (4) we obtain

$$\phi^i_{n-1} = \varphi^i_n + 2\pi(i+1)\, pitch_n \cdot transp_n \cdot \Delta t_{sync} - 2\pi \cdot \frac{f^i_{n-1} \cdot transp_{n-1} + f^i_n \cdot transp_n}{2} \cdot \Delta t + 2\pi(i+1)\Delta t \cdot pitch_n \cdot (transp_n - 1)$$

6.2. Spectral shape concatenation

In order to avoid spectral shape discontinuities at the segment boundaries we make use of the EpR model. First we estimate the EpR of the boundary frames. The left EpR then is stretched using the resonance frequencies as mapping points ($F0_{left}$ *to* $F0_{right}$, $F1_{left}$ *to* $F1_{right}$,…) and it is subtracted from the right EpR. The differential envelope obtained accounts for the spectral shape differences between the two joint frames.

For each one of the transition frames at the left of the boundary, a frequency mapping function is obtained from the interpolation between the above resonance frequency mapping and the identity mapping (y=x) with a factor *1-SSIntp* which stands for the distance from the frame to the boundary (*SSIntp* is 0 at the beginning of the interpolation zone and 1 at the boundary). This mapping is applied to each transition frame spectral shape and finally the differential envelope (weighted by *SSIntp*) is added to it (see Figure 7).

Notice that the spectral shape interpolation is spread along several frames in a similar way to the phase concatenation, but with the addition of the spectrum stretching.

7. CONCLUSIONS

The singing synthesizer we have presented in this paper has proven to comprise suitable methods and techniques for the generation of synthetic vocal tracks.

The SPP has brought a significant improvement in comparison to our previous analysis / synthesis technique [7] based on sinusoidal plus residual decomposition [6]. SPP avoids the tricky sinusoids / noise decomposition and preserves accurately and consistently the voice quality of the original singer. The preservation of the voice quality is a very valued feature when it comes to create databases of singers with vocal disorders such as hoarseness or breathiness.

Even tough the system is in a stage in which a synthetic voice is distinguishable from a real voice the quality is good enough to use it for backing vocals with no synthetic voice perception at all. Anyway we believe we will not have to wait that long to hear synthesized lead vocals that sound indistinguishable from a human singer.

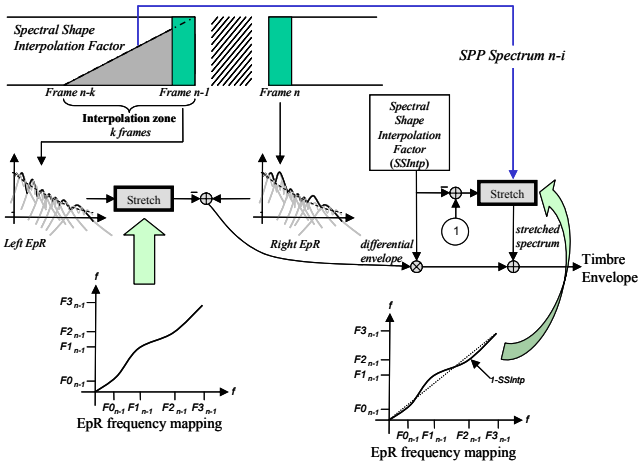

Figure 7: *The sample transformation*

8. REFERENCES

[1] Cook, R., *Toward the Perfect Audio Morph? Singing Voice Synthesis and Processing*, Proceedings of the Digital Audio Effects Workshop DAFX, Barcelona 1998.

[2] Bonada, J., Loscos, A., Cano, P., and Serra, X, *Spectral Approach to the Modeling of the Singing Voice*, Proceedigns of the 111th AES Convention, New York, USA 2001.

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

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

[5] Laroche, J. and Dolson, M., *Improved phase-vocoder. Time-Scale Modification of Audio*, IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 323-332, May 1999.

[6] Serra, X, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*., PhD thesis, CCRMA, Dept. of Music, Standord University, 1989.

[7] Bonada, J., Celma, O., Loscos, A., Ortolà, J., and Serra, X., *Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models*, Proceedings of the International Computer Music Conference, Havana, Cuba 2001.

[8] Bonada, J., *Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio*, Proceedings of the International Computer Music Conference, Berlin, Germany 2000.