A survey of methods for the synthesis of the singing voice

Vesa Siivola
43036b

Presentation for the course S-89.155, sound synthesis

November 19, 2002

Abstract

In this paper, different methods for synthesis of the singing voice are studied. Of the models presented here, physical models have probably the strongest theoretical background. The spectral models try to generate a natural sounding voice by mimicking the spectrum of a real voice. In this paper, spectral modeling with vocoders, formant synthesizers, formant wave function synthesizers and frequency modulation are presented. The most natural sounding source for human singing voice is a human singer. One approach, common in speech synthesizers, is to make a database of real voices and use these to synthesize the voice. Since building a big enough database is not feasible, the system must be able to perform transformations to fill the missing slots in the database. In this paper, concatenative synthesis with pitch synchronous overlap add, analysis by synthesis/overlap add and excitation plus resonances are studied.

Contents

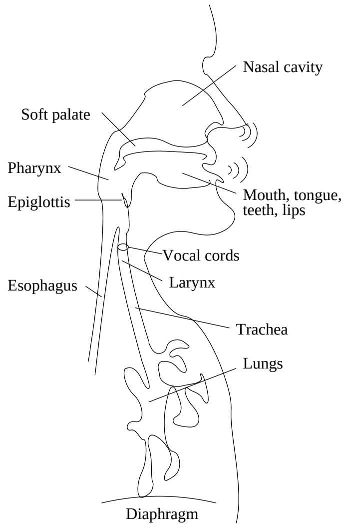


Figure 1: *Vocal tract.*

1 Background

1.1 Vocal tract

The vocal tract as a musical instrument is both flexible and restricted. Humans can produce a wide range of different sounds, but there are even more sounds that humans can most certainly not produce. From modeling point of view, we have to find the right balance of flexibility and restrictions for our model. If we restrict the model to a few well-modeled and natural sounds, we easily notice, that we miss a lot of things that a real singer could do. In the other hand, if we make out model very flexible, we find that it is hard to keep the synthesis natural, since the wide range of controls allow as to produce also unnatural sounds.

The schematic drawing of vocal tract is presented in figure 1. The energy source for voice production is lungs. When muscles reduce the volume of the lungs, the increased air pressure escapes through the glottis. If the produced sound is voiced, the singer uses the muscles of the vocal cords to position the the cords so that they vibrate. The tension of the muscles controls the frequency of the cords' vibration. This vibration leads to a triangle-like pressure at glottis shown in figure 2. The corresponding spectrum has a harmonic structure. When the pressure wave encounters the other parts of the vocal tract, the resonances of the tract shape the spectrum so, that there are a clearly more energetic parts in the spectrum. These spectral peaks are called

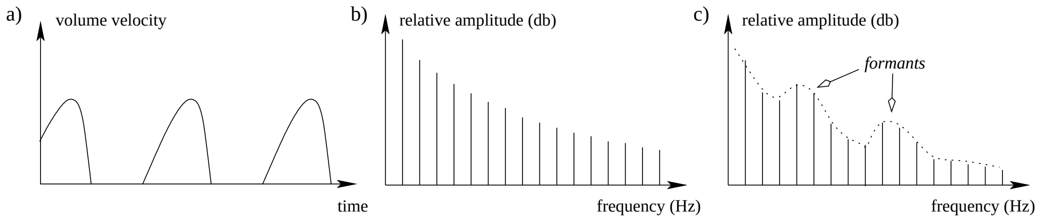


Figure 2: *Glottis pulse. a) air volume velocity, b) spectrum of a glottis pulse, c) spectrum of glottis pulse, that has gone through the vocal tract*

formants and are essential for recognizing voiced phonemes. For unvoiced sound, instead of the glottal source, the excitation is the turbulence of air. The location of this turbulence can be controlled by narrowing the vocal track in appropriate places (throat, teeth, lips). In this manner the singer can produce fricative phonemes (/h/,/s/,/f/, etc). For plosive phonemes (/k/,/t/,/p/, etc) the singer closes a part of a vocal tract. When the closure is opened, the burst of air is heard as a consonant. [Siivola, 1999]

Even though the above description makes modeling of a human voice sound simple, this is not the case. The vocal tract can not be modeled with a simple linear acoustic tube, since the every part of the tube is soft tissue, which absorbs the vibration differently. The shape of the glottis pulse can vary greatly depending of the vocal effort and the other source variations like hoarseness or softness of the voice. The control mechanism for the shape of the vocal tract is very intricate. For example the tongue is almost completely made of muscle and is capable of precise, complicated and elaborate movements.

1.2 Interesting phenomena in a singer's voice

Although there is not a clear consensus, how vibrato is produced (pure frequency modulation of the source or modulation of vocal tract resonances), it can be quite well be modeled by modulating source frequency. Vibrato is a strong expressive tool for stressing important parts and making the sound more lively.

A singer can not hold his voice completely steady. *Jitter* is the variation of the pitch of the voice at rates higher than the typical vibrato frequency. It is caused by random neural firing in the auditory feedback chain. The gradual sifting of tuning is called *drift*. It is more controllable than jitter.

Since vocal tract has mass, it is slow. Pitch change does not occur instantenously. Instead, singer changes pitch at the end of the preceding note

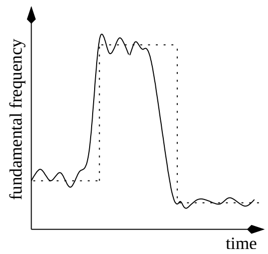


Figure 3: *The pitch of a singer's voice. The solid line is the actual pitch and the dotted line the theoretical target pitch. Sinusoidal change during the steady states is vibrato, the unsinusoidal part jitter. The pitch of the note does not change instantenously, but the move towards correct pitch is executed at the end of the previous note so that the new note starts at the correct pitch.*

so that the new note will start at the correct pitch (see figure 3).

The singer's effort can not be modeled by simply scaling the amplitude. This only sounds like the singer was further away. When singer sings loudly, the power in the high end of spectrum rises significantly more than in the low end. This can be modeled by varying the tilt of the spectral envelope. This effect is audible also in long notes, where the energy of the signal is more or less constant, but the singer effort grows towards the end of the signal.

Since the intelligibility of the words in a song is not as important as in speech, it is sometimes compromised to produce a more powerful singing voice. For female singers, the first formant of a vowel is often below the fundamental frequency. Especially classical female singers often move the first formant to coincide with the fundamental and the second formant to coincide with a multiple of the fundamental. This creates strong resonances to some of the harmonics of the note, creating a much more powerful sound. The corresponding effect for male solo singers is a fixed resonance at around 2700 Hz. This colors the voice so that it is easily distinguished from for example backing choir. This is called "the singer's formant". [Sundberg, 1987]

If a singer would sing through a song with a completely neutral voice, the result would sound unpleasant, like the singer would not be at all interested in the singing. To add some expressiveness to the song, many factors have to be considered. The *phrasal structure* of the music must be found. The start of the phrase has often accelerando and the end a final ritardando. Especially the final note is usually lengthened. In addition to this, the harmonic structure of the music often demands an emphasis on certain notes. The significance of the *harmonic change* has to be found [Sundberg, 1991], so that emphasis (vibrato, preceding crescendo, ritardando) can be put to right places.

Singer can emphasize notes also with special techniques like *marcato*, which is produced by clear overshoot of the target frequency at the beginning of the note before returning to correct pitch. Last, syllable and note coordination has to be carefully observed to produce natural sounding singing.

1.3 A short history of singing voice synthesis

The synthesis of singing voice draws influences from well researched fields, the coding and synthesis of speech and synthesis of musical instruments. Speech synthesis has been studied already from the end of 18th century [Kempelen, 1791], when Wolfgang von Kempelen built a mechanical synthesizer. The first well-known electric synthesizer was created Leon Theremin in 1920s. The first singing voice synthesizer was the vocoder (VOice CODER) by Homer Dudley [Dudley, 1939]. By tweaking parameters of the channel vocoder, it was possible to produce singing voice. In another early effort in late 1950s, Kelly and Lochbaum made an acoustic tube model of the vocal tract, which could produce singing voice [Kelly and Lochbaum, 1962].

There are many ways to synthesize singing voice. One is to make a physical model the voice production of a singer. By modeling the glottal source and turbulances and resonances of the vocal tract, a model capable of producing singing voice can be constructed. [Titze and Story, 1993]

There is also a wide range of pseudo-physical models. Common to all of these models is that they decompose the model to a source and the vocal tract. The classical example of this is the linear prediction, where the vocal tract resonances are modeled as poles of a filter and the remaining residual can be considered as source. The problem with the approach is that modifying the filter does not produce expected results, since there is much more than only the glottal excitation left in the source signal. This is partly due to the nonlinearities of the vocal tract, which this kind of linear method can not model. [Cook, 1996]

It would probably be more correct to classify linear prediction as a spectral method. Common to the spectral methods is that they try to model the spectrum produced by the singer. Vocoders divide the speech spectrum into channels, for which gain and source parameters are approximated. By transforming these parameters, a singing voice can be produced. Also formant synthesizers bear a close resemblance to linear prediction [Sundberg, 1991], [Pabon, 1993]. There is usually a choice of two different source types, unvoiced (white noise) and voiced (sequence of pulses). The vocal tract is modeled as a set of formant filters.

The formant wave functions [Rodet, 1984c] and the frequency modulation [Chowning, 1989] are a bit different spectral methods. The formant wave

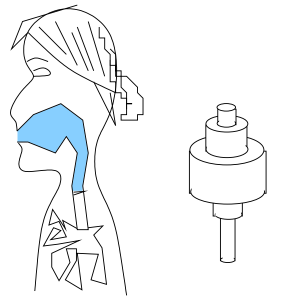


Figure 4: *Modeling of the vocal tract. Here the light blue region (the vocal tract from glottis onwards) is modeled by acoustic tubes in series.*

functions model the impulse response of the formants in time domain. Each of these formant wave functions can be excited at the required fundamental frequency to produce the singing voice.

Frequency modulation synthesis has been popular in musical synthesizers from 80's, but it is not widely used in synthesizing singing. Frequency modulation synthesis is based on the idea, that when a carrier wave is modulated by a sinusoid, a harmonic spectrum can be produced. If we put a suitably scaled and modulated carried wave close to formants of the singing, we can approximate the spectrum of a singing voice.

Another way to model singing voice is to learn directly from recorded samples and then modify these samples to produce required singing voice. Pitch Synchronous Overlap Add (PSOLA) method does this in time domain [Schnell et al., 2000a], [Dan et al., 1993]. The problem is that it is not trivial to transform the waveform in time domain so that it remains natural sounding. Since the human hearing is largely based on frequencies, we can also learn the short-time spectrum of the singing voice and transform this. Here, it is easier to make natural sounding transformations to the signal. The problem with this model is that the the way the control parameters affect the sound is not always very straightforward. [Macon, 1996],[Ortolà, 2001]

2 Methods

2.1 Physical models

A straightforward way of building a speech synthesizer is to model the actual voice organs of a human. In this model, the vocal tract is usually decomposed to source and actual vocal tract. The kinematic and biomechanical properties of glottis were modeled in [Titze, 1984]. The complete modeling of the vocal tract would require solving the differential wave equation

$$\Delta^2 \psi = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial^2 t^2} \tag{1}$$

By the simplifying assumption, that the walls of the vocal tract are hard (ψ is zero), we still end up with a very complex equation. Simply modeling the vocal tract as a long pipe, we can solve the differential equation by d'Alembert's method [Weisstein, 2002]. This gives us a solution of the form

$$\psi(x - vt, x + vt) = f(x + vt) + g(x - vt), \tag{2}$$

where f and g can be any functions. They represent two waveforms traveling to opposite directions. Of course, to model the vocal tract as a single long pipe is an oversimplification. Instead we can model the vocal tract as consisting of acoustic tube of different radius connected together (see figure 4). Solving the equation for this kind of system [Cook, 2002], gives us equations similar to impedance of an electric circuit, where part of the waveform goes through the joint and part of it is reflected back. This is called a digital waveguide model.

In [Cook, 1993], a complete software package for generating singing voice is presented. The glottal pulses can be generated from hand made templates and the vocal tract is modeled with a digital waveguide model. The Singing Physical Articulatory Synthesis Model (SPASM) is a graphical user interface for generating different vocal tract shapes and glottis pulses. The nasal tract is modeled as a parallel acoustic tube joining the main vocal tract. Also more subtle effects, like radiation through throat wall, are modeled. The LECTOR software package can convert text and scores to parameters. The Singer synthesizer software can read this score of parameters and generate the singing voice.

2.2 Spectral models

2.2.1 Vocoders

Vocoding is the oldest technique presented here. The original vocoder was developed at Bell labs [Dudley, 1939]. The idea is to divide the frequency spectrum into bands, and code the energy of the each band. The bands are usually placed so, that they are auditorily equally broad (see figure 5).

In addition to the power of the signal, the source analyzer has to decide, if the source is voiced or unvoiced and in case of the voiced excitation, also the fundamental frequency. The speech signal can be transformed by tweaking the control parameters: The gain parameters control the spectral envelope

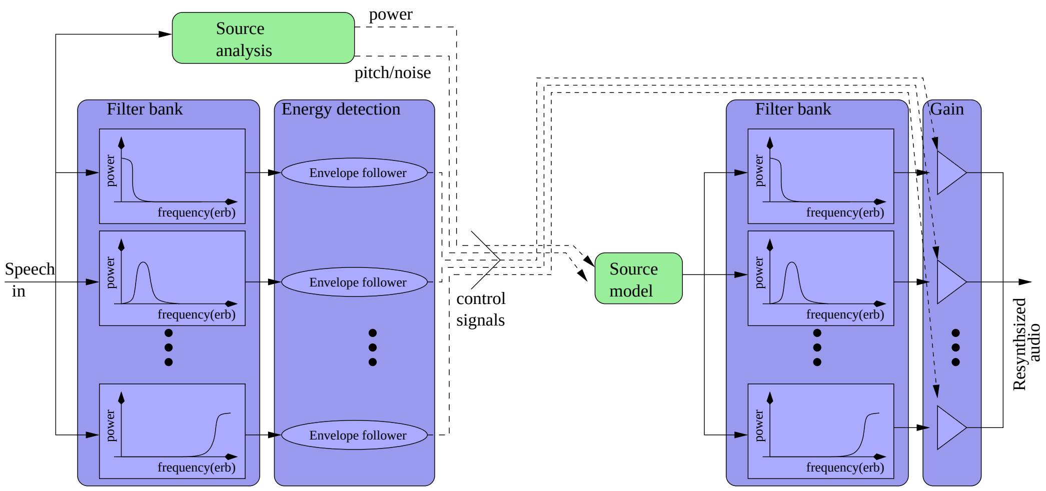


Figure 5: *Vocoder. The left side shows the analysis part and the right side the synthesis part. The control signals are drawn with dashed line.*

(formants, type of fricatives) and the source parameters the pitch and the amount of voicing. To actually synthesize singing voice with the vocoder, all of the control parameters must be specified. This is not a easy and producing even close to natural sounding singing is hard. [Cook, 2002]

The vocoder has often been used to create special effects, like talking cows or singing guitars. Here, the excitation signal is taken from other source than the gain parameters. This is called cross-vocoding.

2.2.2 Formant synthesizers

The formant synthesizers are created by exciting a vocal tract model (filter bank with formant resonances). There is usually a different sources for voiced (glottis pulse) and unvoiced (air turbulence) sounds. The vowel sounds are created by tuning the formant resonances and the pitch can be change simply by changing the excitation rate. The MUSSE DIG synthesizer [Sundberg, 1989] is a typical formant synthesizer (see figure 6). The musical expressiveness parameters are generated automatically by the RULSYS-system. The formant synthesizers bear a strong resemblance to vocoders.

2.2.3 Formant wave function synthesis

Formant wave function synthesis (FOF, Forme d'Onde Formatique) is based on the idea, that we can model the impulse responses of each formant individually. We can then sum the output of each formant wave function to

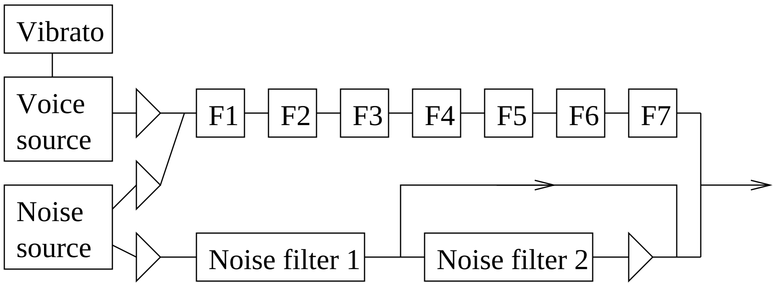


Figure 6: *The MUSSE DIG singing voice synthesizer. The noise source can be driven through the whole vocal tract or through the noise filters, which simulate the noise source being closer to mouth, for example at lips. After [Berndtsson and Sundberg, 1993].*

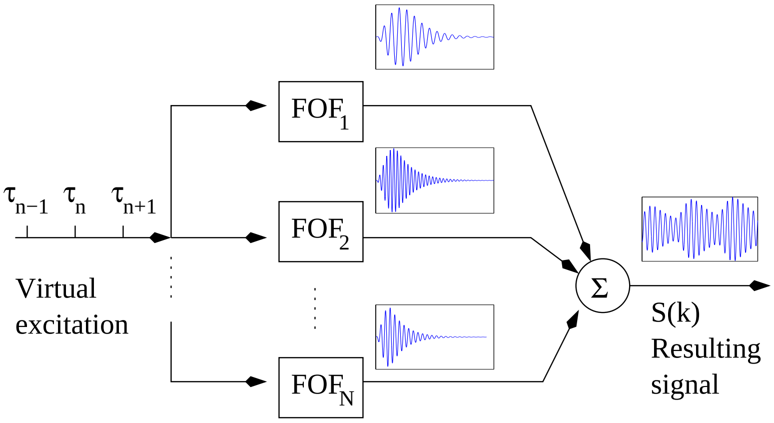


Figure 7: *The formant wave function synthesizer, after [Bennett and Rodet, 1989].*

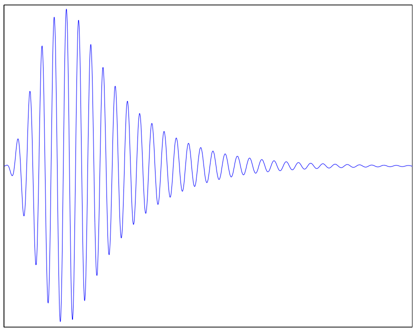


Figure 8: *A formant wave function*

form the singing voice (see figure 7). Each of the formant wave functions are exited once in the start of the pitch period.

A reasonable model for a formant wave function is

$$s(k) = 0, \quad k < 0 \tag{3}$$

$$s(k) = \frac{1}{2}\Big((1 - cos[\beta k]) \cdot e^{-\alpha k} sin[\omega k + \Phi]\Big), \quad 0 \leq k \leq \frac{\pi}{\beta} \tag{4}$$

$$s(k) = e^{-\alpha k} sin[\omega k + \Phi]), \quad k > \frac{\pi}{\beta} \tag{5}$$

Here, the cosine term of equation 4 controls the speed of onset of the formant and the exponential term (in both eqs. 4, 5) controls the decay. An example wave formant function is shown in figure 8.

As formant wave function synthesizer is a time domain tool, the accuracy of calculations is not that critical (for example, there will be no unstable poles because of rounding errors, as in linear prediction). Formant wave functions are computationally cheap. If necessary, they can be reduced to simple table lookups. [Rodet, 1984c]

There is an implementation of formant wave synthesis called CHANT [Rodet, 1984a]. This software has quite extensive control of the voice, including controls for example for vibrato, hoarseness, vocal effort and jitter. A higher level language to control the synthesizer, called FORMES, was also created [Rodet, 1984b].

2.2.4 Frequency modulation (FM)

Frequency modulation attempts to generate a spectrum that closely approximates the spectrum of the singing voice by modulating a carrier wave by a sinusoidal:

$$s(t) = A \sin[2\pi f_c t + I \sin(2\pi f_m t)] \tag{6}$$

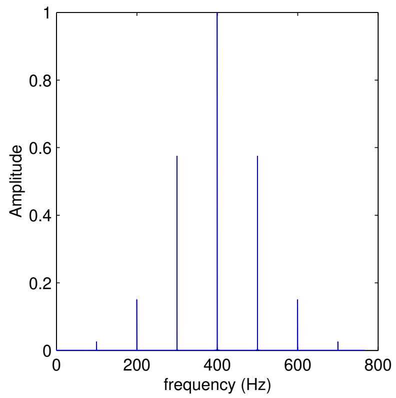


Figure 9: *Frequency modulation: a 400 Hz carrier signal modulated by 100 Hz sinusoidal. Modulation index is 1. Using a few carriers with the same modulation signal, one can construct a voice-like spectrum, where the formants lie at the carrier frequencies.*

Here, f_c is the carrier frequency and f_m the modulating frequency. The ratio of these frequencies determines the relative spacing between the component frequencies in the modulated carrier signal. If $f_c = Nf_m$, the modulated signal will be a harmonic series, where the fundamental f_0 is equal to f_m. The modulation index I determines, how much energy is transferred from the carrier to the sidebands, thus controlling also, how many sidebands are generated.

In [Chowning, 1989], the formants of the singing voice are generated with three parallel FM pairs. The f_m (corresponding to fundamental frequency) is the same for all of the FM oscillators and the carrier frequency f_c determines the center of the formant. The modulation index I controls the width of a formant. A FM synthesizer capable of synthesizing three formants can be described with

$$\begin{aligned} s(t) &= A_1^{0.5}\sin[2\pi f_{c1}t + I_1\sin(2\pi f_0 t)] \\ &\quad + A_2^{1.5}\sin[2\pi f_{c2}t + I_2\sin(2\pi f_0 t)] \\ &\quad + A_3^{2}\sin[2\pi f_{c3}t + I_3\sin(2\pi f_0 t)] \qquad (7) \end{aligned}$$

The singer effort is modeled elegantly with this model. As speaker effort increases, the high frequency content of the signal increases. This is modeled by taking a higher power of the amplitude of the higher formants. To make the synthesized voice sound more natural, jitter and vibrato must be added to the f_0.

Although frequency modulation synthesis is widely used in music synthesis, it has not gained much popularity in singing voice synthesis.

2.3 Concatenative synthesis

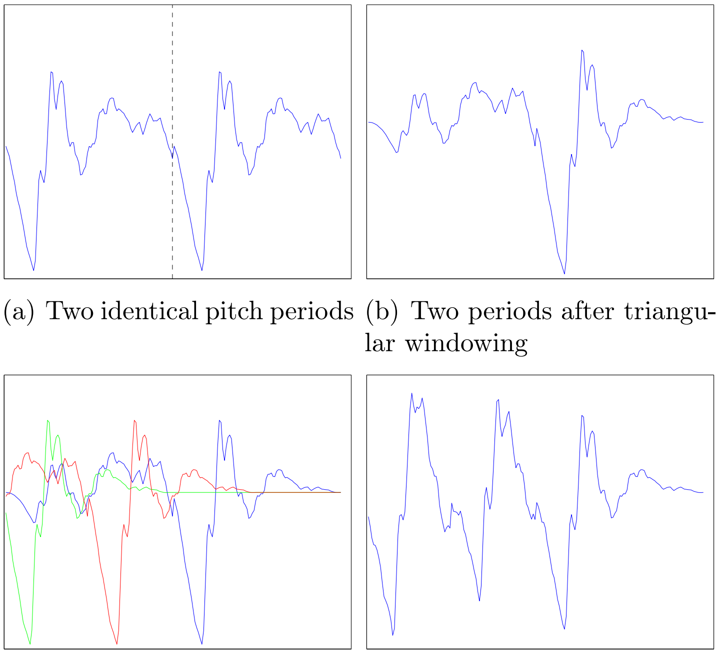
Concatenative synthesis is based on the idea, that if we have large enough database (with any kind of sound we could desire), we could just pick the relevant sounds from the database and put them one after another and thus get the desired synthesized sound. Of course, in practice this it is impossible to get or even store such a database. Instead, we can find a way to change a sound of the database into another sound, so that the new sound will also sound natural. If we have flexible enough transforms, we can cope with a reasonably sized database. This method comes from speech synthesis researchers.

Example: We need to produce a sung “a” with pitch $g\flat$ sung loudly (forte). The preceding phoneme we have already synthesized was “u”. From the database we find vowel “a” with pitch $a\sharp$ and loudness mezzoforte which was preceded by “v”. Since “v” and “u” produce similar coarticulation effects, the pitch change from $a\sharp$ to $g\flat$ is quite small and the loudness is similar, we can expect to produce quite natural sounding vowel.

We can store the exact waveforms in the database, as in Pitch Synchronous Overlap Add (PSOLA) method. Another way is to parameterize the stored samples so, that the future transformations will be easier. Sinusoidal coding tries to represent the signal as a sum of few sinusoidal signals. This kind of sparse coding is leads also to relatively easy ways to transform the signal (pitch change, time-scale modification).

2.3.1 Pitch-Synchronous Overlap Add (PSOLA)

PSOLA is a time domain method. It decomposes the signal into elementary waveforms. Each waveform corresponds to one pitch period, and when these waveforms are summed with overlap add, the original signal is reconstructed (see figure 10). Time shift can be accomplished by repeating or skipping pitch periods. Frequency modification is done by changing the timing between the elementary wave functions. The PSOLA works only for voiced parts of the signal, since there has to be a fundamental period for the model to work. In [Schnell et al., 2000a], the unvoiced parts of the signal were stored as having a constant pitch period. During resynthesis, the unvoiced samples were concatenated in random order, thus reducing the possibility of accidental correlations, which could result in reverberating, tonal or flanging voice. The PSOLA method copes well with small transformations, but for bigger changes, the voice quality suffers. The method bears some resemblance to



(a) Two identical pitch periods (b) Two periods after triangular windowing

(c) Three windowed signals (d) Resulting synthesis signal

Figure 10: *The reconstruction of a voice from one pitch period. In the subfigure a), the original stored pitch period is plotted twice. In preparation for overlap-add, the double pitch period is windowed with a triangular window (subfigure b). To produce a higher fundamental frequency, the windowed signals are placed closer to each other (subfigure c). When the overlapping windows are summed, we get the synthesized speech signal in subfigure d).*

wavetable synthesis methods.

2.3.2 Analysis By Synthesis/Overlap Add (ABS/OLA)

The ABS/OLA method divides the signal to overlapping windows. For each window, the algorithm tries to find a number of constant-frequency sinusoidals, that best model the window. The sinusoids are found by iterative analysis-by-synthesis method, which tries to minimize the squared error on each iteration.

The signal can thus be presented by

$$x[n] = \sigma[n] \sum_{k=0}^{K-1} w_s[n - kN_s] s_k[n - kN_s]. \tag{8}$$

Here, σ controls the envelope of the produced sound, w_s is the window func-

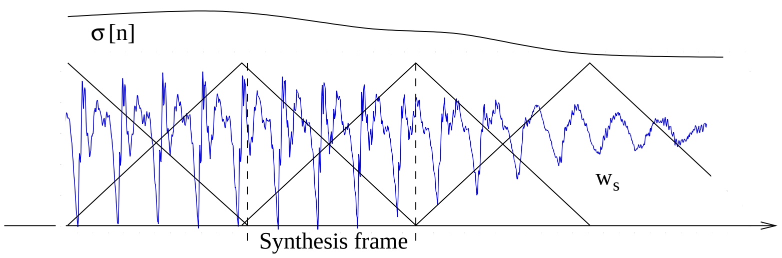


Figure 11: *ABS/OLA overlap add window*

tion (usually triangle) and s_k is the sum of sinusoids belonging to this window. A sketch of this is presented in figure 11. The sum of of sinusoids s_k can written as

$$s_k[n] = \sum_{l=0}^{L-1} A_l * k \cos(w_l kn + \phi_l^k). \tag{9}$$

The actual resynthesis of the signal is straightforward, since it only involves inverse Fourier transform and overlap-add. The complexity of the algorithm is in finding the sinusoidals, which can be done off-line when we construct our database. The actual synthesis is computationally relatively light.

Frequency-Scale and Time-Scale Modification

A simple way to expand the length of a sample would be to make each window longer. Also the gain envelope $\sigma[n]$ would have to be stretched. Changing frequency scale would just mean scaling the frequencies of sinusoidals. Changing of the pitch would require removing the spectral characteristics of vocal tract, changing the frequencies and rescaling with the vocal tract spectral envelope.

Unfortunately, since in practice, the sinusoidal components are not completely harmonically related, these methods lead to phase coherence problems. For voiced parts, the phase coherence can be restored by tweaking the phase of the sinusoidals. For unvoiced parts, the problem is a bit harder. To avoid tonal or reverberant sounding synthesis, the phase of the signal in the unvoiced parts should be randomized. [Macon, 1996]

Simple rescaling of sinusoidals with formant tract spectral envelope can bring noisy parts under formants, resulting in poor voice quality. This can be avoided by synthesizing the signal from resampled spectrum of the stored signal. [Macon, 1996]

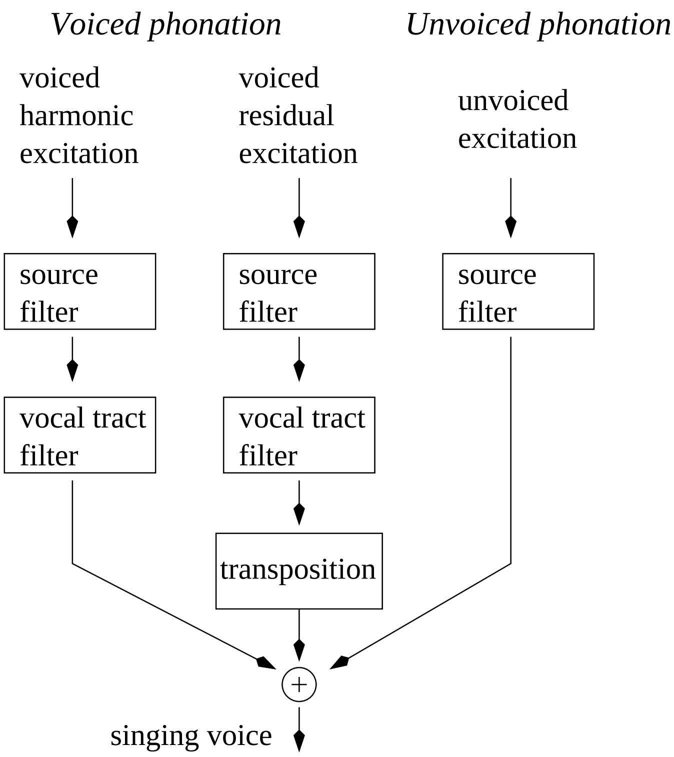


Figure 12: *Excitation plus Resonances synthesizer. After [Bonada et al., 2002].*

2.3.3 Excitation plus Resonances (EpR)

In EpR model starts with Spectral Modeling Synthesis (SMS). The SMS differs a bit from the sinusoidal coding presented in previous section. Instead of trying to find optimal coding for each analysis window separately, it tries to optimize the problem over all windows so can we track, how each sinusoidal changes. The change can be change of frequency, dying away or appearing. The model also keeps track of the residual signal $e(t)$.

$$s(t) = \sum_{r=1}^{R} A_r(t) \cos[\phi_r(t)] + e(t) \tag{10}$$

The vocal tract response is estimated for stored samples. The sinusoidals and residual are inverse filtered before they are stored into database. This helps the resynthesis, since now the source and vocal tract parameters are separated and can be transformed individually. The source filter models the singer effort, since it is responsable for the overall tilt of the spectrum.

The synthesis signal path of the actual EpR synthesizer is in three parallel channels (see figure 12). The voiced harmonic excitation are the sinusoidals.

The voiced residual excitation goes through same filters as the main excitation, but has to be scaled in frequency domain, because it still holds some characteristics of the actual sound it was learned from. The unvoiced sounds are directly stored, they are only first inverse filtered with the estimated source filter.

Both the vocal tract filter resonances, the residual excitation and the unvoiced excitation are learned from a large number of samples and stored into a database. In [Ortolà, 2001], there are samples for different attack and release (normal, sharp, soft, sexy), different dynamics (soft, normal, loud), and different transitions (staccato, legato, glissando, portamento), different size of mouth opening, different singing styles (normal, flat, happy, sad, jazzy, drunk) and different vibratos.

3 Summary

In this paper, several methods for singing voice synthesis were presented. As a subjective evaluation, it seems that as the synthesis model comes more flexible, the tasks the model is able to do become more restricted. As the number of control parameter increases, it becomes increasingly difficult to create natural voice. The concatenative synthesis methods are the most restricted ones, since they can only produce, what they can transform from the example samples in their database. On the other hand, they are the only methods capable of producing even tolerably natural sounding singing voice with lyrics. The spectral methods can produce quite convincing singing voice, if no lyrics are required. The Queen of Night's coloratura (from the Magic Flute) sung by CHANT-synthesizer [Potard and Rodet, 1984] was by far the most convincingly natural example, but to get the synthesizer to sound that natural, a lot of tweaking by hand was required. While the physical models rest on the strongest theoretical foundation, they still have a long way to go to produce convincing singing voice.

Of the presented methods, only vocoding has been widely used by entertainment industry. It is used to create special effects, like robot voices or making the speaker sound like dragon. The movie Farinelli [Corbiau, 1994] tells a story about baroque era castrato singers. Since the voices of castrato singers have many properties, that no living singer can match, the required vocal parts had to be created by computer. The most of the work was done by cross-vocoding a tenor singer's and a soprano singer's voice with a phase vocoder [Depalle et al., 1995].

Since human voice is still the best source for singing voice, it is hard to see that a computer could gain a significant foothold in generating natural singing voice. The biggest role that computer plays, is to enhance and correct the errors of the real singer. The possibility to create subtly unnatural sounds has a much wider use for special effects and also for artistic expression. In these areas, the use of computers can be expected to grow.

Finally, there are some examples of singing voice synthesis in the world wide web. Listening to the examples can provide insight into current state-of-art technology. [Potard and Rodet, 1984], [Macon et al., a], [Macon et al., b], [Schnell et al., 2000b], [KAE Labs, 2002], [Bell Labs, 1940]

References

[Bell Labs, 1940] Bell Labs (1940). Voder. http://ptolemy.eecs.berkeley.edu/~eal/audio/vocoder.html.

[Bennett and Rodet, 1989] Bennett, G. and Rodet, X. (1989). Synthesis of the singing voice. In Mathews, M. and Pierce, J., editors, *Current Directions in Computer Music Research*, chapter 4, pages pp. 19–44. MIT Press, Cambridge, Massachusetts.

[Berndtsson and Sundberg, 1993] Berndtsson, G. and Sundberg, J. (1993). The MUSSE DIG singing synthesis. In *Proceedings of the Stockholm Music Acoustics Conference*.

[Bonada et al., 2002] Bonada, J., Celma, Ó., Loscos, À., Ortolà, J., Serra, X., Yoshioka, Y., Kayama, H., Hisaminato, Y., and Kenmochi, H. (2002). Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models. In *Proc. Int. Computer Music Conf*, pages 139–146, Havanna, Cuba.

[Chowning, 1989] Chowning, J. (1989). Frequency modulation synthesis of the singing voice. In Mathews, M. and Pierce, J., editors, *Current Directions in Computer Music Research*, chapter 6, pages pp. 57–63. MIT Press, Cambridge, Massachusetts.

[Cook, 1993] Cook, P. (1993). SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. *Computer Music Journal*, 17(1):pp. 30–44.

[Cook, 1996] Cook, P. (1996). Singing voice synthesis: History, current work, and future directions. *Computer Music Journal*, 20(3):pp. 38–46.

[Cook, 2002] Cook, P. (2002). *Real Sound Synthesis for Interactive Applications*. A. K. Peters.

[Corbiau, 1994] Corbiau, G. (1994). Farinelli. movie.

[Dan et al., 1993] Dan, T., Mukherjee, B., and Datta, A. (1993). Temporal approach for synthesis of singing. In *Proceedings of the Stockholm Music Acoustics Conference*.

[Depalle et al., 1995] Depalle, P., Garcia, G., and Rodet, X. (1995). The recreation of a castrato voice, Farinelli's voice. In *IEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*.

[Dudley, 1939] Dudley, H. (1939). The vocoder. Bell Laboratories record.

[KAE Labs, 2002] KAE Labs (2002). Vocalwriter. http://kaelabs.com/.

[Kelly and Lochbaum, 1962] Kelly, J. and Lochbaum, C. (1962). Speech synthesis. In *Proceedings of the Fourth International Congress on Acoustics*, pages pp. 1–4.

[Kempelen, 1791] Kempelen, W. (1791). *Le mechanisme de parole, suivi de la description d'une machine parlante*. J.V. Degen, Vienna.

[Macon, 1996] Macon, M. (1996). *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology. http://www.ece.ogi.edu/∼macon/publications/.

[Macon et al., a] Macon, M., Clements, M., Jensen-Link, L., Oliverio, J., and George, E. Flinger – Festival Singer. http://cslu.cse.ogi.edu/tts/flinger/.

[Macon et al., b] Macon, M., Clements, M., Jensen-Link, L., Oliverio, J., and George, E. The LYRICOS singing voice synthesis system. http://cslu.cse.ogi.edu/tts/research/sing/sing.html.

[Ortolà, 2001] Ortolà, J. (2001). Musical and phonetic controls in a singing voice synthesizer. Master's thesis, Polytechnics University of Valencia.

[Pabon, 1993] Pabon, P. (1993). A real time singing voice synthesizer. In *Proceedings of the Stockholm Music Acoustics Conference*.

[Potard and Rodet, 1984] Potard, Y. and Rodet, X. (1984). Synthesis of the aria the queen of the night from Mozart's opera The magic flute. http://www.ircam.fr/equipes/analyse-synthese/reine.html.

[Rodet, 1984a] Rodet, X. (1984a). The Chant project: From the synthesis fo the singing voice to synthesis in general. *Computer Music Journal*, 8(3):pp. 15–31.

[Rodet, 1984b] Rodet, X. (1984b). FORMES: Composition and scheduling of processes. *Computer Music Journal*, 8(3):pp. 32–50.

[Rodet, 1984c] Rodet, X. (1984c). Time-domain formant-wave-function synthesis. *Computer Music Journal*, 8(3):pp. 9–14.

[Schnell et al., 2000a] Schnell, N., Peeters, G., Lemouton, S., Manoury, P., and Rodet, X. (2000a). Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA). In *Proc. Int. Computer Music Conf.*, pages pp. 102–108.

[Schnell et al., 2000b] Schnell, N., Peeters, G., Lemouton, S., Manoury, P., and Rodet, X. (2000b). Virtual choir sound examples using PSOLA synthesis. http://www.ircam.fr/anasyn/Bastille-K/index.html.

[Siivola, 1999] Siivola, V. (1999). An adaptive method to achieve speaker independence in a speech recognition system. Master's thesis, Helsinki University of Technology.

[Sundberg, 1987] Sundberg, J. (1987). *The Science of the Singing Voice*. Northern Illianois University Press.

[Sundberg, 1989] Sundberg, J. (1989). Synthesis of singing by rule. In Mathews, M. and Pierce, J., editors, *Current Directions in Computer Music Research*, chapter 5, pages pp. 45–55. MIT Press, Cambridge, Massachusetts.

[Sundberg, 1991] Sundberg, J. (1991). Synthesizing singing. In Poli, G. D., Picialli, A., and Roads, C., editors, *Representations of Musical Signals*, chapter 9, pages pp. 299–320. MIT Press.

[Titze, 1984] Titze, I. (1984). Parametrization of the glottal area, glottal flow and vocal fold contact area. *Journal of Acostic Society of America*, 75(2).

[Titze and Story, 1993] Titze, I. and Story, B. (1993). The Iowa singing synthesis. In *Proceedings of the Stockholm Music Acoustics Conference*.

[Weisstein, 2002] Weisstein, E. (2002). Eric Weisstein's word of mathematics. http://mathworld.wolfram.com/WaveEquation.html.