Toward the Perfect Audio Morph?
Singing Voice Synthesis and Processing

Perry R. Cook
Princeton University Computer Science Department (also jointly in Music)
prc@cs.princeton.edu http://www.cs.princeton.edu/~prc

Abstract

This paper reviews the popular methods and models used for the synthesis of the singing voice, discussing strengths and weaknesses of each technique. Then a brief review is given of research on cross-modal visual/auditory perception of the human voice. The paper concludes with comments related to the singing synthesis systems discussed, addressing multi-modal perception, audio morphing, and the categorical perception of sound.

1 Introduction

The human voice is the most ubiquitous, flexible, and general of acoustic instruments. We all have one, yet only a few of us learn to "play" it with proficiency as a musical instrument. Most functions of this instrument we take for granted, but huge regions of our brains are dedicated to controlling and perceiving the sounds made by it. Even those people that never learn to use it musically still are able to perform amazing feats of imitation and flexibility with their voice. The voice can exact independent control across a broad range of pitch, amplitude, brightness, harmonicity, noise amount, and spectral shape.

The voice cannot be taken apart and studied like most other instruments. We cannot "build" versions with small variations in the parameters to observe the effects. We cannot try different materials and structures, such as the violin maker/ player might do with different woods, varnishes, bracing structures, strings, bows, and rosins. The true subtleties of a fine singing voice must be studied, "in vivo," if at all, and only with the graceful cooperation of the owner/builder/ player/instrument (all one in the same).

The attraction of composers to the human voice instrument has a rich, long history. In modern times, computer music composers have time and again been attracted to vocal sounds and processing. Part of this is due to the legacy of computer music tools, with many of them arising from the great speech labs of the world. But there is more at work than the mere availability of tools.

Many historical electronic audio effects devices intentionally mimic the human voice (vocoders). Others sound vocal in some sense, by the sheer nature of one particular feature such as a resonance that can be swept independently of the source sound parameters (the wah-wah pedal). One might expect that we could look to vocal analysis/synthesis techniques to give us ideas for new digital audio effects, perhaps informing us as to how to create the "perfect audio morph."

This paper will survey models, methods, and systems for the analysis, synthesis, and processing of the human voice. It focuses on singing synthesis and voice-related tools which have found use in computer music composition. The positive and negative aspects of each system or model will be noted. Finally, areas of research and perception of vocal sounds (and images) will be discussed.

2 Singing Voice Methods, Models, and Systems

The voice has traditionally been viewed as a linear source/filter system. That is, there are one or more sources of sound, and one or more filters which shape the spectrum of those sound sources. By moving various articulators, we change the ways the sources and filters behave.

The voice source can be characterized as a periodic source corresponding to the oscillating vocal folds, or a non-periodic source corresponding to turbulent noise, or a mixture of these. The voice system filter properties are controlled by the shape of the vocal tract.

The spectrum of the voice is characterized by resonant peaks called formants. Figure 1 is a spectrum corresponding to the vocal vowel / i / (as in beet), showing harmonics of the voice source outlining the peaks and valleys of the vocal tract filter response.

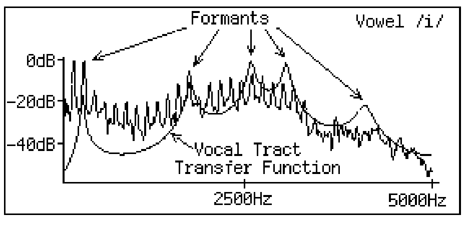

Figure 1. Voice spectrum for the vowel /i/ (as in beet), showing harmonics and formant peaks.

The location and shapes of formant resonances are strong perceptual cues that we use to identify vowels and consonants. The most successful systems capable of generating, recognizing, or flexibly modifying speech-like sounds, have allowed flexible manipulation of the resonant peaks of the spectrum, and of source parameters (voice pitch, noise level, etc.).

2.1 Spectral Subband Vocoders

From the early legacy of speech signal processing came the powerful and flexible signal processing techniques known as the spectral subband vocoders (VOice CODERs). In the channel vocoder [1] and phase vocoder [2][3], the spectrum is broken into sections called subbands, and the information in each subband is analyzed. The analyzed parameters are then stored or transmitted for reconstruction at another time or physical site. The parametric data representing the information in each subband can be manipulated, yielding transformations such as pitch or time shifting, or spectral shaping.

The channel vocoder models only the time-varying amplitude within each subband, and typically uses between 10 and 30 subbands to cover the entire audible spectrum. Figure 2 shows a block diagram of a channel vocoder. This architecture yields well to implementation in analog circuitry, and a number of analog hardware devices were produced and sold as musical instrument processing devices in the 1970-80's. One attraction of these devices, as with other source/filter models of the voice, is that the source can be replaced with arbitrary sounds, resulting in talking cows, singing guitars, etc. This is called cross-synthesis.

Since the channel vocoder explicitly makes an assumption that the signal being modeled is a single human voice, it does not generalize to arbitrary sounds, and fails horribly when the source parameters deviate from expected harmonicity, reasonable voice pitch range, etc.

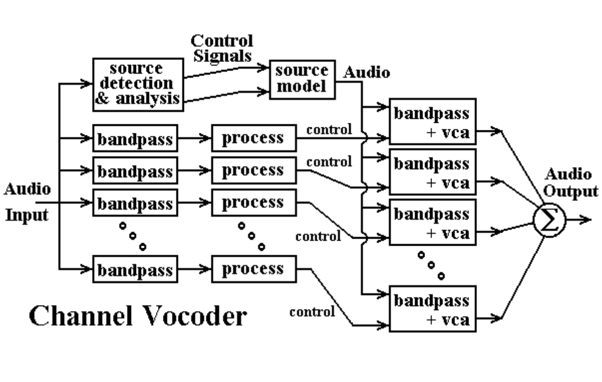

Figure 2. Block diagram of a channel vocoder

The phase vocoder calculates and maintains both instantaneous magnitude and phase, and is implemented using the Fast Discrete Fourier Transform. Many subbands (sinusoidal DFT bins) are typically used (on the order of hundreds to thousands). Unlike the channel vocoder, the phase vocoder does not perform an explicit source/filter decomposition, and there is no parametric model of the source. The phase vocoder does not strictly assume that the signal is speech, and thus can generalize to other sounds. For this reason, the phase vocoder has found extensive use in computer music composition.

By the nature of FFT processing; segmenting the signal in blocks of many samples, then analyzing it into an equal number of subbands, the phase vocoder does nothing to make sonic data more parametric. For composition, data reduction or compression is not necessarily a goal. However, sound analysis systems which in some way make data parametric often make for good composition systems, allowing manipulation of a relative few parameters rather than thousands of numbers per second. For anything other than simple time and spectrum stretching, more processing must be done on the raw spectral data yielded by the phase vocoder. We will discuss this further in the section on spectral modeling systems..

2.2 Linear Prediction

Linear Predictive Coding (LPC) [4], as shown in Figure 3, involves forming a digital filter that predicts the next time sample from a linear combination of a few previous samples. An error signal is yielded which, if fed back through the time-varying prediction filter, will yield exactly the original signal. The filter models linear correlations in the signal, which correspond to spectral features such as formants. The error signal models the input to the formant filter, and typically is periodic and impulsive for voiced speech, and noise-like for unvoiced speech. The error signal can be parametrically coded and resynthesized, or modified before resynthesis.

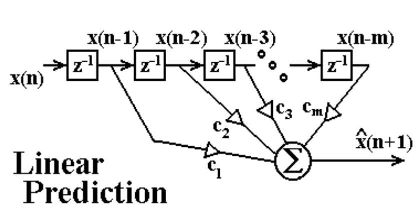

Figure 3. A linear predictive digital filter.

The success of LPC in representing speech signals is largely due to the similarity between the source/filter decomposition yielded by the mathematics of linear prediction, and the source/filter model of the human vocal tract. The introduction of LPC revolutionized speech technology, and had a great impact on musical composition as well [5][6][7]. The power of LPC as a compositional tool stems from the ability to modify the parameters before resynthesis. As with the channel vocoder, the source can be replaced with arbitrary sounds, allowing for cross synthesis.

In LPC, however, all spectral properties are modeled in the filter. In actuality the voice has multiple possible sources of non-linear behavior, including source-tract coupling, non-linear wall vibration losses, and aerodynamic effects. Due to these deviations from the ideal source-filter model, the result of analysis/modification/resynthesis using LPC or a subband channel vocoder often sounds artificial. One further problem with LPC is that the least-squares method of determining the optimal filter coefficients causes the designed filter to match well at peaks, but less well at spectral valleys (see Figure 1).

2.3 Frequency Modulation

Frequency Modulation (FM) involves modulating the frequency of one oscillator (the carrier) with the output of another (the modulator) to create a spread spectrum consisting of sidebands surrounding the carrier frequency. For FM sound synthesis, both carrier and modulator operate in the audio frequency range. The most easily described scheme for FM sound synthesis is that in which both the carrier and modulator oscillators generate sinusoidal waveforms. In this case, sinusoidal sideband frequencies are generated at the carrier frequency, the carrier frequency plus and minus the modulation frequency, the carrier frequency plus and minus two times the modulation frequency, and so on. As a rough rule of thumb, the number of significant sidebands is equal to the index of modulation (the ratio of carrier frequency deviation to modulation frequency) minus two.

FM sound synthesis as introduced by Chowning [8][7], proved successful for the synthesis of a variety of sounds, including the synthesis of singing. By controlling the amount of modulation, and using multiple carrier/modulator pairs, spectra of somewhat arbitrary shape can be constructed. This technique proved extremely efficient for digital synthesis, yet sufficiently flexible for music composition. In vocal modeling, carriers placed near formant locations in the spectrum are modulated by a common modulator oscillator operating at the voice fundamental frequency. Figure 4 shows a block diagram of a simple FM voice synthesizer.

In order to generate a harmonic voice spectrum using FM synthesis, the carrier frequencies must be integer multiples of the fundamental modulator frequency. For this reason, it is impossible to generate vocal sounds which smoothly vary arbitrarily from vowel to vowel, or from pitch to pitch on a single vowel. Also, there is no closed-form analysis technique for identifying FM parameters to yield an identity resynthesis of an arbitrary sound.

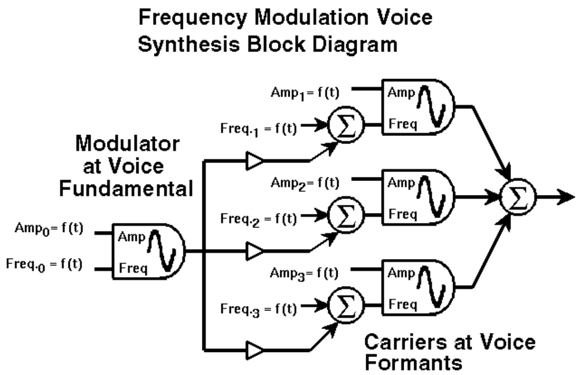

Figure 4. FM voice synthesis block diagram.

2.4 FOFs

Formant Wave Functions (FOFs in French) represent time-domain waveform models of the impulse responses of individual formants [9]. These are characterized as a sinusoid at the formant center frequency with an amplitude which rises rapidly upon excitation and decays exponentially. By describing a spectral region as a windowed sinusoidal oscillation in the time domain, FOFs can be viewed as a special type of wavelet. The control parameters define the center frequency and bandwidth of the formant being modeled, and the rate at which the FOFs are generated and added determines the fundamental frequency of the sound. Figure 5 shows the process of adding FOFs to create a voice waveform.

The synthesis system for controlling FOFs was dubbed CHANT, and has found application in general music synthesis [10] as well as synthesis of the singing voice [7]. The parametric FOF description of spectral features allows for continuous manipulation of those features. As such, the CHANT system provides a convenient dual description of sonic features in terms of either the time or frequency domain.

The basic FOF parameters, however, might not be the most convenient for composers. Also, FOFs do not directly allow for cross-synthesis to be performed between two sounds, as is easily accomplished using the channel vocoder or LPC.

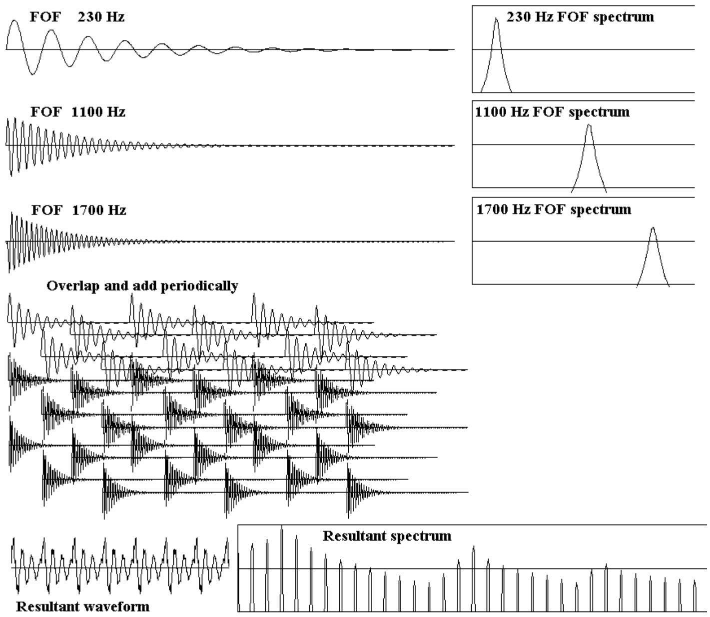

Figure 5. Three FOFs (top), added and overlapped at a periodic rate, generate a voice waveform and spectrum.

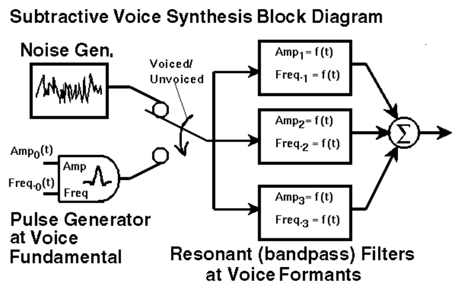

Figure 6. Formant synthesizer block diagram.

2.5 Formant Filter Models

Second order resonant filters can be used to model formants directly [11][12]. An attractive feature of formant synthesizers is that Fourier or LPC analysis can be used to automatically extract formant frequencies, bandwidths, and source parameters from recorded speech. Computer music composers have used formant vocal models for composition [7].

The Speech Transmission Laboratory of the Swedish Royal Institute of Technology created the MUSSE DIG (MUsic and Singing Synthesis Equipment, DIGital version) [13]. This system has been used in singing synthesis [14], for studying performance synthesis-by-rule [7], and has been adapted for real-time control [15].

Formant filters provide parametric control over what might be the most "speechlike" spectral feature, however, the assumption is still one of a strictly linear model. Speech and singing researcher Johan Sundberg has often been heard to say "none of us has ever seen a formant," implying that there is much more to the voice than a simple linear model.

2.6 Sinusoidal Models

As noted in Section 2.1, simply performing a Fourier transform on speech data does not yield a parameterization which is useful beyond simple pitch and time manipulations. Sinusoidal speech modeling [16] uses Fourier analysis to locate and track individual sinusoidal partials in the voice signal. Individual trajectories (tracks) of sinusoidal amplitude, frequency, and phase as a function of time are extracted from the time varying peaks in a series of Short Time Fourier Transforms (STFT). To help define tracks, heuristics regarding physical systems and the voice in particular are used, such as the fact that a sinusoid should not appear, disappear, or change frequency or phase instantaneously.

The sinusoids can be resynthesized from the track parameters, after modification or coding, by additive synthesis. Noise can be treated as rapidly varying sinusoids, or explicitly as a non-sinusoidal, stochastic component [17]. The technique of modeling the deterministic (sinusoidal) and stochastic (noise) components separately is called Spectral Modeling Synthesis, and has found use in music composition. Figure 7 shows a deterministic/ stochastic decomposition of a sound wave.

2.7 Acoustic Tube/Physical Models

Acoustic tube models simulate the vocal tract transfer function by solving the one dimensional wave equation inside a smoothly varying tube. The one dimensional approximation is justified by noting

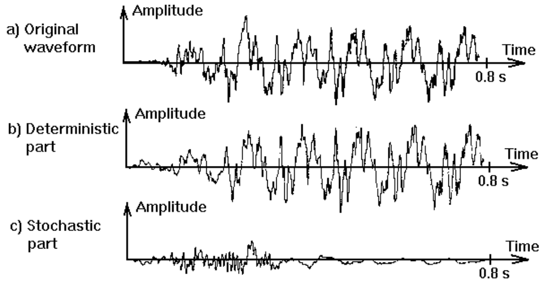

Figure 7. A sound waveform (upper), the purely deterministic part as modeled by sinusoids (center), and the stochastic residual (lower). (courtesy X. Serra)

that the length of the vocal tract is significantly larger than any width dimension, and thus the longitudinal modes dominate the resonance structure up to about 4000 Hz. Modal standing waves in an acoustic tube correspond to the formants. Early speech modeling work at Bell Labs included the acoustic tube model of Kelly and Lochbaum [19]. The basic Kelly-Lochbaum model critically samples space and time by approximating the smooth vocal tract tube with cylindrical segments equal in length to the distance traveled by a sound wave in one time sample. Figure 8 shows a smooth acoustic tube, the sampled version of that, and a ladder filter model of the sampled tube, with Kelly-Lochbaum scattering matrix operations at the junctions of adjacent tube sections.

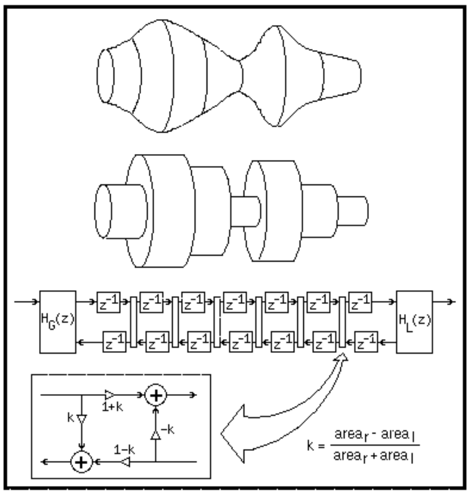

Figure 8. Smooth acoustic tube, a sampled version, and a waveguide ladder filter simulation.

The SPASM and Singer [19] systems are based on a Kelly-Lochbaum physical model of the vocal tract filter, motivated by the waveguide formulation [20]. The SPASM model is a direct descendent of the Kelly-Lochbaum model, but with many enhancements, such as a nasal tract, modeling of radiation through the throat wall, various steady and pulsed noise sources [21], and real-time controls. The SPASM/Singer model also adds natural inertial parameters to the basic acoustic tube model, yielding interpolations from shape to shape automatically.

Maeda's [22] acoustic tube model numerically integrates the wave equation using the rectangular method in space, and the trapezoidal rule in time. Wall losses are also modeled, and an articulatory layer of control modifies the basic tube shape from higher-order descriptions like tongue and jaw position. Carre's [23] model is based on Distinctive Regions (DR) arising from sensitivity analysis, noting that movements in particular regions of the vocal tract affect formant frequencies more than movements in others. Liljencrants [24] investigated an undersampled acoustic tube model and derived rules for modifying the shape without adding unnaturally to the energy contained within the vocal tract. Acoustics researchers in Helsinki [25] have used fractional sample interpolation and truncated conical tube segments to derive an improved version of the Kelly-Lochbaum model.

2.8 Model Variants and Other Systems

Pabon [26] has constructed a singing synthesizer, with real-time formant control via spectrogram-like displays called phonetograms, and source waveform synthesis using FOF-like controls. Titze and Story [27] have produced a super-computer tenor called "Pavarobotti," which is used for studying many aspects of the voice including advanced physical models of normal and pathological vocal folds.

Ken Lomax at Oxford University, and the Lyricos project at Georgia Tech have constructed systems based on spectral templates, using spectral modeling techniques. Lomax [28] has tackled the difficult problem of characterizing, archiving, and resynthesizing the unique voices and singing styles of famous singers. The Lyricos [29] project dealt with synthesis of arbitrary segments of singing from a small set of example sounds. One additional spectral-template-based project involved the cross synthesis of analyzed soprano and counter-tenor singing, to create a virtual castrato singer for the movie "Farinelli (Il Castrato)" [30].

3 Spectral and Physical Models

Synthesis models can be loosely broken 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. Both physical and spectral models have merit, and one or another might be more suitable given a specific goal and set of computational resources.

Of the models and techniques discussed above, the spectrally-based models include FM, FOFs, phase and channel vocoders, and sinusoidal models. Acoustic tube models are physically-based, while formant synthesizers are spectral models, but could be classified as pseudo-physical because of the source/filter decomposition. LPC can be interpreted in three ways; as a least-squares linear prediction of the time domain waveform, as a least squares matching process on the spectrum, and as a source-filter decomposition. Therefore, LPC is both a spectral and pseudo-physical model, but not strictly a physical model because wave variables are not propagated directly in the simulation, and no articulation parameters go into the basic model. LPC can be mapped to a filter related to the acoustic tube model [31], thus creating a bridge between the spectral and physical camps.

The main attraction of physical models is that the control parameters are those that a human uses to control his/her own vocal system. As such, some intuition can be brought into the design and composition processes. Another motivation is that time-varying model parameters can be generated by the model itself, if the model is constructed so that it sufficiently matches the physical system.

Disadvantages of physical models are that the number of control parameters can be large, and while some parameters might have intuitive significance for humans (jaw drop), others might not (specific muscles controlling the vocal folds). Further, parameters often interact in non-obvious ways. Finally, in general there exist no exact methods for analysis/resynthesis using physical models.

Spectral models, by virtue of being based on frequency domain features, are undeniably related to some aspects of the human perceptual mechanism. The cochlea as frequency transformer, the tonotopic mapping of the auditory cortex, etc. all closely relate to the Fourier Transform. Indeed frequency domain representations have proven the best spaces so far in which to talk about "audio morphing."

But vocoders, FFTs, and time-varying sinusoidal tracks do not actually match any known structures in the auditory system. There are no sinusoids in the human brain, no Hanning windows, and no buffers in convenient lengths of powers of two. There are proponents of the (still hotly debated) "motor theory of speech perception," which asserts that we use articulatory gestures directly to perceive speech sounds. The parameter spaces yielded by spectrally based systems are not necessarily the most natural ones for composition, manipulation, recognition, compression, or the study of perception. Much must be added to the parameters of vocoders, or sinusoidal models, to make them truly useful. Much of this mapping and parameterization is still unknown, but it remains an exciting area for future research.

4 Faces, Lips, and Voice Perception

An interesting area of research and artistic endeavor involves facial animation coupled with voice synthesis. This is of interest perceptually because humans use a significant amount of lip reading in understanding speech and singing. The two modalities compliment each other, with information that is difficult to discern using only one sense often disambiguated by the other. However, interesting work has been done to investigate cases where the two sensory modalities disagree [32]. Work has been done by Massaro [33] and others [34], employing facial animation to study coupling of visual and auditory information in human speech understanding.

Musically, we know that the face of the singer can carry even more information about the meaning of music than the actual text being sung [35], further motivating the combination of facial animation with singing synthesis. Work with simultaneous analysis of audio and facial video has allowed signal processing to be performed on the speech sound, in conjunction with image processing on the video, yielding convincing faces saying things they never actually said in real life [36].

5 Morphing, Genus, and Categorical Perception

This final section will briefly address the following:

- Is the voice itself, or voice analysis/synthesis/ modeling, the place to look for the perfect audio morph?
- What is an audio morph anyway? Is it appropriate to take a term, concept, etc. from one sensory modality (or media sub-discipline) and carry it directly into another?
- Is there something more interesting to do compositionally with voice modeling systems beyond pitch shifting, time shifting, and cross synthesis?

Even the most cursory search of the speech literature yields many papers on vowel spaces, and on the categorical perception of vowels. Exciting recent work by Kuhl [37] and others has used cross-cultural infant studies to investigate the process of learning and acquisition of vowel templates (these templates to be used later for adult speech production and recognition). It is clear from this work and many vowel perception studies is that there is not a smooth continuum in any known vowel space, and the perception of vowels tends to be categorical. This of course makes sense; if we are to understand speech at all, we must "round" sounds to a nearby vowel, otherwise all sounds which do not match exactly our internal templates will not be perceived as speech.

The issue of categorical perception goes to one of the fundamental issues of vocal modeling, vocal processing, the use of voice-based computer models to process arbitrary sounds, and computer music composition in general. Researchers designing models, systems, and tools for audio synthesis and manipulation often make claims such as "we will be able to create entirely new instruments, not subject to the restrictions of existing acoustic instruments," or "we will be able to synthesize an instrument that is halfway between instrument-A and instrument-B." This author has, of course, made such claims. Let's briefly examine those issues; entirely new sounds, sounds halfway between two known sounds, and sounds which smoothly vary from one known sound to another.

One learns in working with synthesis tools that it is very difficult, if not impossible, to create "entirely new sounds." Sounds tend to sound "like" something. This again seems to be an artifact of the necessities of our linguistic processing systems. It is more likely that we will generate something that listeners describe as "a trumpet-like thing with too much amplitude modulation," than to generate something that illicits a description of "wow, that's half-way between an oboe and a trumpet."

Even synthesizing a "perfect" vocal spectrum, but omitting random and periodic pitch deviation, causes the perception to stray from that of a voice. Likewise, those instruments that are more often described as "singing," or "voicelike"; the violin, the Theremin(oVox), some wind instruments; have those attributes related to the fine pitch and amplitude control that are most typical of the human voice.

The term "audio morph," has been applied to transitions between two instruments, between two vocal vowels, between two musical phrases, and even between the compositional styles of two composers. None of these are really equivalent to a morph in the graphics domain. There is a known fact in topology that the only morphs that are mathematically well posed are those that take place between two objects of the same geometric genus. This doesn't stop anyone from trying to morph a sphere (or human head) into a coffee cup, with possibly interesting artistic results. But geometric morphs that do not obey this basic rule are not possible to pose or compute uniquely, and generally don't tend to work well perceptually.

In the audio domain, we have an intuitive (or experienced) feel that it would be easier to move smoothly from clarinet to trumpet to flute to voice, than it would be from piano to voice to snare drum to duck-quack. Indeed, the clarinet, trumpet, flute, and voice share much in common: a non-linear periodic oscillator driven by breath pressure, resonator structure, components of harmonicity and noise, etc. But there are fundamental differences as well: the voice has independent pitch and resonance control, and the "reeds" are different for each of those wind instruments (air for the flute, inwardly beating wood for the clarinet, outwardly beating lips for the trumpet, etc.).

There are some fundamental physical attributes of musical instruments which seem almost the parallel of the topological genus. There are also spectral features and attributes that we can use to say that two sounds are more or less similar. But there are also perceptual categories, groups, and boundaries which challenge the notion of an ideal audio morph. The author feels that the profound linguistic nature of human perception, new knowledge of physical models of sound producing objects, and new studies of timbre in general, should all motivate the questions to be asked in searching for audio morphs and new digital audio effects.

6 Conclusions

There are many ways to analyze, synthesize, and process vocal sounds. Systems intended for speech coding and compression have had a great influence on computer music synthesis and composition. Spectral models, physical models, and others all have their place, if nothing other than to pose interesting questions about sound, sound sources, and perception.

References

[1] Dudley, H. 1939, "The Vocoder," *Bell Laboratories Record*, December.

[2] Moorer, A. 1978. "The Use of the Phase Vocoder in Computer Music Applications." *Journal of the Audio Engineering Society*, 26 (1/2), pp. 42-45.

[3] Dolson, M. 1986, "The Phase Vocoder: A Tutorial," *Computer Music Journal*, 10 (4), pp. 14 -27.

[4] Atal, B. 1970. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." *Journal of the Acoustical Society of America* 47.65(A).

[5] Moorer, A. 1979, "The Use of Linear Prediction of Speech in Computer Music Applications," *Journal of the Audio Engineering Society* 27(3): pp. 134-140.

[6] Steiglitz, K. and P. Lansky 1981. "Synthesis of Timbral Families by Warped Linear Prediction." *Computer Music Journal* 5(3): pp. 45-49.

[7] Mathews, M. and J. Pierce. 1989. *Some Current Directions in Computer Music Research*. Cambridge MA.: MIT Press: pp. 57-63.

[8] Chowning, J. 1981, "Computer Synthesis of the Singing Voice," in *Research Aspects on Singing*, KTH, Stockholm, Sweden, pp. 4-13.

[9] Rodet, X. 1984. "Time-Domain Formant-Wave-Function Synthesis," *Computer Music Journal* 8 (3), pp. 9-14.

[10] Rodet, X., Potard, Y., and J.B. Barrière 1984. "The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General." *Computer Music Journal* 8(3), pp. 15-31.

[11] Rabiner, L. 1968. “Digital Formant Synthesizer” *Journal of the Acoustical Society of America* 43(4), pp. 822-828.

[12] Klatt, D. 1980. “Software for a Cascade/Parallel Formant Synthesizer,” *Journal of the Acoustical Society of America* 67(3), pp. 971-995.

[13] Carlson, G. and L. Neovius 1990. "Implementations of Synthesis Models for Speech and Singing," *STL-Quarterly Progress and Status Report*, KTH, Stockholm, Sweden, 2-3: pp. 63-67.

[14] Zera, J., Gauffin, J., and Sundberg, J. 1984. “Synthesis of Selected VCV-Syllables in Singing,” *Proc. International Computer Music Conference*, IRCAM, Paris, pp. 83-86.

[15] Carlson, G., Ternström, S., Sundberg, J. and T. Ungvary 1991. “A New Digital System for Singing Synthesis Allowing Expressive Control.” *Proc. of the International Computer Music Conference*, Montreal, pp. 315-318.

[16] McAulay, R. and T. Quatieri. 1986. "Speech Analysis/Synthesis Based on a Sinusoidal Representation." *IEEE Trans. Acoust. Speech and Sig. Proc.* ASSP-34(4): pp. 744-754.

[17] Serra, X. and J. Smith 1990. "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition." *Computer Music Journal* 14 (4), pp. 12-24.

[18] Kelly, J., and C. Lochbaum. 1962. "Speech Synthesis." *Proc . Fourth Intern. Congr. Acoust*. Paper G42: pp. 1-4.

[19] Cook, P. 1992. "SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software Synthesis System,” Computer Music Journal, 17: 1, pp. 30-44.

[20] Smith, J. 1987. *Musical Applications of Digital Waveguides.* Stanford University Center For Computer Research in Music and Acoustics. Report STAN-M-39.

[21] Chafe, C. 1990. "Pulsed Noise in Self-Sustained Oscillations of Musical Instruments." *Proc. IEEE Int. Conf. on Acoust. Speech and Sig. Proc.* Albuquerque, NM. 2(S2): pp. 1157-1160.

[22] Maeda, S. 1982. “A Digital Simulation Method of the Vocal Tract System.” *Speech Communication* 1: pp. 199-299.

[23] Carre, R. 1992. “Distinctive Regions in Acoustic Tubes.” *Journal d’Acoustique*, 5(141), pp. 141-159.

[24] Liljencrants, J. 1985. *Speech Synthesis With a Reflection-Type Line Analog*, DS Dissertation, Speech Communication and Music Acoustics, KTH, Stockholm, Sweden.

[25] Välimäki, V. and M. Karjalainen 1994. “Improving the Kelly-Lochbaum Vocal Tract Model Using Conical Tube Sections and Fractional Delay Filtering Techniques.” *Proc. 1994 Int. Conf. on Spoken Language Processing*, Yokohama Japan, pp. 18-22.

[26] Pabon, P. 1993, "A Real-Time Singing Voice Synthesizer," *Stockholm Music Acoustics Conference*, KTH, Stockholm, Sweden, pp. 288-293.

[27] Story, B. and I. Titze 1995. "Voice Simulation With a Body-Cover Model of the Vocal Folds," *Journal of the Acoustical Society of America* 97 (2), pp. 3416-3431.

[28] Lomax, K. 1996, “The development of a singing synthesizer,” Speech and Computers (SPECOM96), pp. 146-150.

[29] Macon, M., Jensen-Link, L., Oliverio, J., Clements, M. and E. George, 1997 “A Singing Voice Synthesis System Based on Sinsusoidal Modeling,” Proc. ICASSP, Vol. 1, pp. 439-442.

[30] Depalle, P., Garcia, G., and Rodet X. 1994, “A Virtual Castrato (!?)” *Proc. International Computer Music Conference*, Aarhus, pp. 357-360.

[31] Markel, J. and A. Gray, 1976, *Linear Prediction of Speech*, New York, Springer.

[32] McGurk, H, and J. MacDonald. 1976. "Hearing Lips and Seeing Voices." *Nature* 264, pp. 746-748.

[33] Massaro, D. 1987, *Speech Perception by Ear and Eye*, Hillsdale, N.J., Erlbaum Associates.

[34] Hill, D., Pearce, A., and B. Wyvill 1988. "Animating Speech: an Automated Approach Using Speech Synthesized by Rules." *The visual computer*. 3 (5), pp. 277-289.

[35] Scotto Di Carlo, N. and I. Guaitella 1995. “Facial Expressions in Singing.” *Thirteenth International Congress of Phonetic Sciences*, Stockholm, Sweden, pp. 1:226-229.

[36] Bregler, C., Covell, M. and M. Slaney 1997. “Video Rewrite: Driving Visual Speech With Audio,” ACM SIGGRAPH 1997, pp. 353-360.

[37] Kuhl, P., Williams, K., Lacerda, F., Stevens, K. and B. Lindblom, 1992 “Linguistic experience alters perception in infants by 6 months of age,” Science, 255, pp. 606-608.