PITCH AND TIMBRE MANIPULATIONS USING CORTICAL REPRESENTATION OF SOUND

D. N. Zotkin, S. A. Shamma, P. Ru, R. Duraiswami, L. S. Davis

Perceptual Interfaces and Reality Laboratory, UMIACS, University of Maryland, College Park 20742

ABSTRACT

The sound received at the ears is processed by humans using signal-processing that separates the signal along intensity, pitch and timbre dimensions. Conventional Fourier-based signal processing, while endowed with fast algorithms, is unable to easily represent signal along these attributes. In this paper we use a recently proposed cortical representation to represent and manipulate sound. We briefly overview algorithms for obtaining, manipulating and inverting cortical representation of a sound and describe algorithms for manipulating signal pitch and timbre separately. The algorithms are first used to create sound of an instrument between a "guitar" and a "trumpet". Applications to creating maximally separable sounds in auditory user interfaces are discussed.

Partial support of ONR grant N000140110571 is gratefully acknowledged.

1. INTRODUCTION

When a natural source such as a human voice or a musical instrument produces a sound, the resulting acoustic wave is generated by a time-varying excitation pattern of a possibly time-varying channel, and the sound characteristics depend both on the excitation signal and on the production system. The production system (e.g., human vocal tract, the guitar box, or the flute tube) has its own characteristic response; variation of the excitation parameters produces a sound signal that has different frequency components, but still retains perceptual characteristics of the uniqueness of the production instrument (identity of the person, type of instrument – piano, violin, etc.) When one is asked to characterize this sound source using descriptions based on Fourier analysis one discovers that concepts such as frequency and amplitude are insufficient to explain the characteristics of the sound source. Human linguistic descriptions characterize the sound in terms of pitch and timbre.

The perceived sound pitch is closely coupled with its harmonic structure. On the other hand, the timbre of the sound is defined broadly as everything other than the pitch, loudness, and the spatial location of the sound. For example, two musical instruments might have the same pitch if they play the same note, but it is their different timbre that allows us to distinguish between them. Specifically, the spectral envelope in frequency and the spectral envelope variations in time are related to the timbre percept. Most conventional techniques of sound manipulation result in simultaneous changes in both the pitch and timbre and cannot be used to assess the effects of the pitch and timbre dimensions independently. A goal of this paper is the development of controls for independent manipulation of pitch and timbre of a sound source using a *cortical sound representation* that was introduced in [1] and used for assessment of speech intelligibility and for prediction of the cortical response to an arbitrary stimulus. We simulate the multiscale audio representation and processing believed to occur in the primate brain (supported by recent psychophysiological papers [2]), and while our sound decomposition is partially similar to existing pitch and timbre separation and sound morphing algorithms (in particular, MFCC decomposition algorithm in [3], sinusoid plus noise model and effects generated with it in [4], and parametric source models using LPC and physics-based synthesis in [5]), the neuromorphic framework allows to view the processing from a different perspective, supply supporting evidence to justify the procedure performed and tailor it to the way the human nervous system processes auditory information, and extend approach to include decomposition in time domain in addition to frequency.

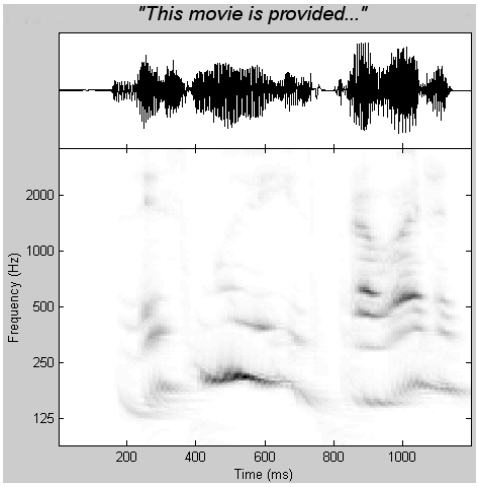

Fig. 1. Example auditory spectrogram for the sentence shown.

We anticipate our algorithms to be applicable in several areas, including musical synthesis, audio user interfaces and sonification. In musical instrument synthesis, synthesizers often use sampled sound that have to be pitch-shifted to produce different notes [5] or generate a new instrument with the perceptual timbre lying in-between two known instruments. Development of advanced auditory user interfaces requires mapping of arbitrary data streams into auditory percepts, and is commonly called “sonification” [6]. Having the ability to create sound objects with controllable pitch and timbre (as well as location and ambience as described earlier in [7]) is necessary for maximal information throughput.

2. THE CORTICAL MODEL

In a complex acoustic environment, sources may simultaneously change their loudness, location, timbre, and pitch. Yet, humans are able to integrate the multitude of cues arriving at their ears, and derive coherent percepts and judgments about each source [8].

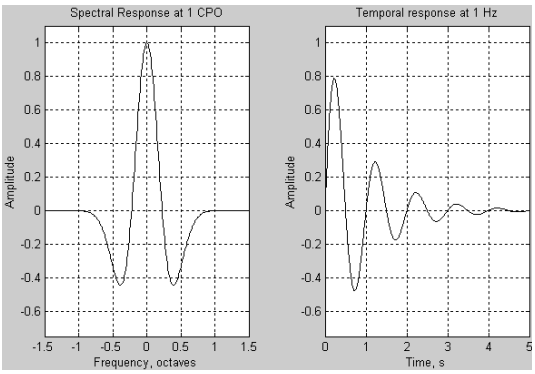

Fig. 2. Tuning curves for the basis (seed) filter for the rate-scale decompostion (scale of 1 cycle per octave, rate of 1 Hz).

The cortical model is a computational model for how the brain is able to obtain these features. Physiological experiments have revealed the elegant multiscale strategy developed in the mammalian auditory system for coding of spatiotemporal characteristics of the sound [2], [9]. The primary auditory cortex (AI), which receives its input from the thalamus, employs a multiscale representation in which the dynamic spectrum is repeatedly represented in AI at various degrees of spectral and temporal resolution. This is accomplished by cells whose responses are selective to a range of spectro-temporal parameters such as the local bandwidth and symmetry of spectral peaks, and their onset and offset transition rates. A mathematical model of the early and central stages of auditory processing in mammals was recently developed and described in [1]. It is a basis for our work and is briefly summarized here.

The first stage of the model is an early auditory stage, which models the transformation of the acoustic signal into an internal neural representation. called the "auditory spectrogram". The second is a central stage, which analyzes the spectrogram to estimate its spectro-temporal features, specifically its spectral and temporal modulations, using a bank of modulation selective filters mimicking those described in the mammalian primary auditory cortex.

The second stage is the auditory spectrogram stage that consists of a sequence of three operations. A frequency analysis is performed first by the mechanical vibrational pattern of a basilar membrane in mammalian cochlea, with different frequencies resonating at different points along the membrane; this stage is simulated by a constant-Q filter-bank wavelet transform operation. Then, the mechanical vibrations are transduced by inner hair cells into electrical pulses in the auditory nerve fibers, which are simulated by half-wave rectification and lowpass filtering of the filter outputs. Finally, lateral inhibition and envelope detection occurs in the anteroventral cochlear nucleus, which is modeled by a spatial derivative across the channel array to sharpen selectivity of filters. Half-wave rectification followed by short-term integration is applied to model the slow adaptation of the central auditory neurons. The resulting time-frequency representation (auditory spectrogram, Figure 1) is invertible through an iterative process. A time-slice of the spectrogram is called the auditory spectrum.

The second analysis stage mimics the action of higher central auditory stages (especially the primary auditory cortex). The findings of a wide variety of neuron spatio-temporal response fields (SRTF) covering a range of frequency and temporal characteristics

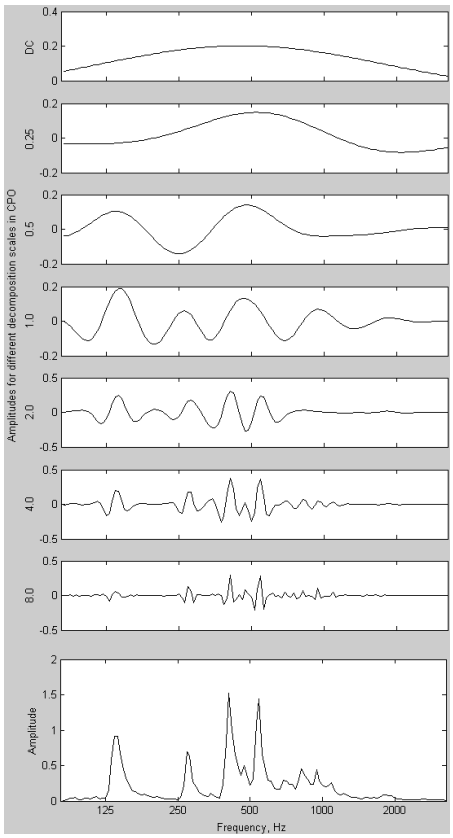

Fig. 3. Sample scale decomposition of the auditory spectrum (bottom plot) using scales from 0.125 to 8.0 CPO (top 7 plots).

[9] suggests that they may as a population perform a multiscale analysis of their input spectral profile. Specifically, the cortical stage estimates the spectral and temporal modulation content of the auditory spectrogram by a bank of modulation selective filters. Each filter is tuned to a combination of spectral and temporal modulation of the incoming signal over a range of temporal modulations, or "rates", varying from 2 to 32 Hz, and spectral resolutions, or "scales", varying from 0.25 to 8 cycles per octave (CPO); filters are centered at different frequencies along the tonotopic axis. Figure 2 shows the spectral and the temporal response of the seed filter tuned for scale 1 CPO and rate 1 Hz; differently tuned filters are obtained by dilation or compression along the spectral and temporal axes. A mathematical formulation of the filter is available in [1]. The filter output is computed by a convolution of its spectro-temporal impulse response (STIR) with the input auditory spectrogram, producing a modified spectrogram. Since the spectral and temporal cross-sections of an STIR are typical of a bandpass impulse response in having alternating excitatory and inhibitory fields, the output is large only if the spectrogram modulations are tuned to the rate, scale, and direction of the STIR. A map of the responses across the filterbank provides a unique characterization of the spectrogram that is sensitive to the spectral shape and dynamics over the entire stimulus (see Figure 3).

Alternatively, the rate-scale analysis can be viewed as the decomposition of a two-dimensional spectrogram using a set of basis

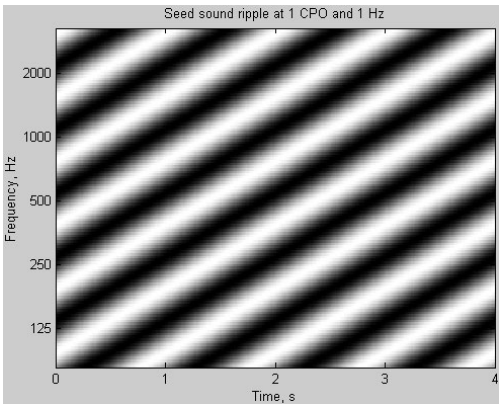

Fig. 4. Sound ripple at scale of 1 CPO and rate of 1 Hz.

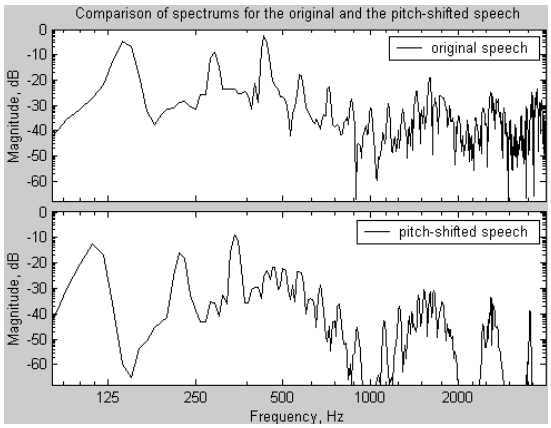

Fig. 5. Spectrum of a speech signal before and after pitch shift. Note that the spectral envelope is filled with new set of harmonics.

functions which are called *sound ripples* and are characterized by their scale and rate. Thus, a ripple with scale 1 CPO and rate 1 Hz has alternating peaks and valleys in the spectrum with 1 CPO periodicity, and the spectrum shifts in time, repeating itself with 1 Hz periodicity (Figure 4); it is a basic seed function for decomposition. All other basic functions are obtained by dilation (compression) of this function in both time and frequency axes. The result of the decomposition of the auditory spectrogram using a basis of sound ripples is a four-dimensional (time, frequency, scale and rate) hypercube of (complex) filter coefficients that can be modified and inverted back to the acoustic signal.

Just as with the forward path, the inversion (or sound reconstruction) consists of an early and central part. The early auditory stage is inverted by an iterative convex projection algorithm that takes the spectrogram as input and reconstructs the acoustic signal that produces the closest spectrogram to a target. The second part of the algorithm is the inversion of the cortical multiscale representation back to a spectrogram. This is critical, since the timbre, pitch, and elevation manipulations are easier to do in the cortical domain. This is a one step inverse filtering operation, followed by a rectification to ensure that the resulting spectrogram is positive.

3. TIMBRE-PRESERVING PITCH MANIPULATIONS

For speech and musical instruments, timbre is conveyed by the envelope of the spectrum, whereas the pitch is mostly conveyed by the harmonic structure, or harmonic peaks. This biologically based analysis is in the spirit of the cepstral analysis used in speech [10], except that the Fourier-like transformation in the auditory system is carried out in a local fashion using kernels of different scales. The cortical decomposition is expressed in the complex domain, with magnitude a measure of the local bandwidth of the spectrum, and phase the local symmetry at each bandwidth. Finally, just as with cepstral coefficients, the spectral envelope varies slowly. In contrast, the harmonic peaks are only visible at high resolution. Consequently, timbre and pitch occupy different regions in the multiscale representation. If X is the auditory spectrum of a given data frame, with length N equal to the number of filters in the cochlear filter-bank, and the decomposition is performed over M scales, then the matrix S of scale decomposition has M rows, one per scale value, and N columns. If further the $1st$ row of S contains the decomposition over the finest scale and the Mth row is the coarsest one, then the components of the S in the upper right triangle (above the diagonal) can be associated with pitch, while the rest of components are the timbre information.

To control pitch and timbre separately, we apply modifications at the appropriate scales, and invert the cortical representation back to the spectrogram. Thus, to shift the pitch while holding the timbre fixed we compute cortical multiscale representation of the whole sound and shift (along the frequency axis) the triangular part of every time-slice of the hypercube that holds the pitch information while keeping timbre information intact and invert the result. To modify the timbre keeping the pitch intact we do the opposite. It is also possible to splice in pitch and timbre information from two speakers, or from a speaker and a musical instrument. The result after inversion back to a sound is a "musical" voice that sings the utterance (or a "talking" musical instrument).

We show one pitch shift example here and refer the interested reader to the web [11], [12] for actual sounds used in this example, and for more samples. We use the above described algorithm to perform timbre-preserving signal pitch shift. The cochlear model has 128 filters with 24 filters per octave, covering $5\frac{1}{3}$ octaves along frequency. The processing is done all the way from the sound wave through the auditory spectrogram to the multiscale representation, the pitch is shifted, and the inversion is carried back to sound wave. In Figure 5, we show the plot of the spectrum of the original and the shifted by 8 channels (one-third of an octave) signal at a given time slice. Pitches of the original and of the modified signals are 140 Hz and 111 Hz, respectively. It can be seen that spectral envelope is preserved and speech formants are kept at their original locations, but a new set of harmonics is introduced.

The algorithm is sufficiently fast to be performed in real-time. To achieve real-time performance, we use simple fast Fourier transform algorithms instead of a cochlear bank filter in both the forward and backward stages of the early auditory processing stage, which eliminates the need for an iterative inversion process. We keep the phase information of the original signal and patch it to the set of amplitudes for final inversion to ensure smooth evolution of phases between frames to prevent artifacts in the synthesis. Additionally, we do not compute the full cortical representation of the sound but perform only scale decomposition of the auditory spectrogram because shifts are done in the frequency axis only and can be performed in each time slice of the hypercube independently.

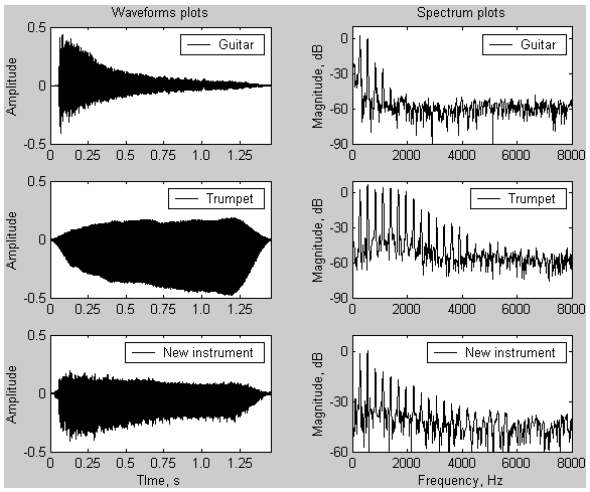

Fig. 6. Wave and spectrum for guitar, trumpet and new instrument.

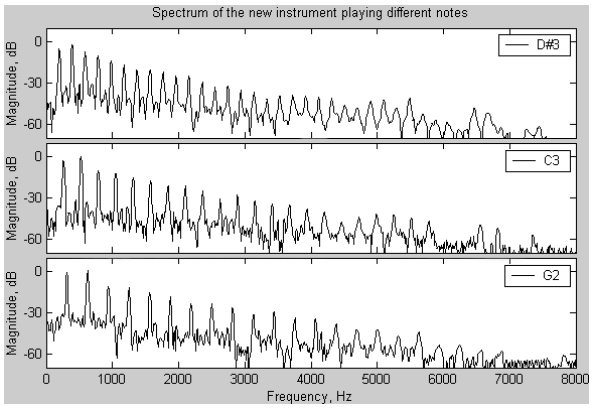

Fig. 7. Spectrum of the new instrument playing D#3, C3 and G2.

4. TIMBRE MANIPULATIONS

Timbre is captured in the multiscale representation both by the spectral envelope and by the signal dynamics. Spectral envelope variations or replacements can be done by modifying the lower right triangle of the multiscale representation of the auditory spectrum, while sound dynamics is captured by the rate decomposition. Selective modifications to enhance or diminish contribution of components of a certain rate can make the sound abrupt or slurred, or create an impression of an anechoic or extremely reverberant environment (see [11] for samples). In musical synthesis, playback rate and onset and decay ratio can be modified while preserving the pitch using shifts along the rate axis.

To show ease of timbre manipulation using the cortical representation, we performed timbre interpolation between two musical instruments to obtain a new in-between synthetic instrument which has the spectral shape and spectral modulation in time (onset and decay ratio) that is in the middle between two original instruments, which we selected to be guitar, $W_g C\#3$, and trumpet, $W_t C\#3$, playing the same note (C#3). Then, the rate-scale decomposition of a short (1.5 seconds) instrument sample was performed and the arithmetic average of the rate-scale representations for two instruments was converted back to the new instrument sound sample $W_n C\#3$. The behavior of the new instrument along the time line is intermediate between two original ones, and the spectrum shape is also an average spectrum of two original instruments (Figure 6).

Then, we used timbre-preserving pitch shift described above to synthesize different notes of the new instrument, using the waveform $W_n C\#3$ obtained in the previous step (third waveform in Figure 6) as an input. Figure 7 shows the spectrum of the new instrument for three different notes – D#3, C3 and G2. It can be seen that the spectral envelope is the same in all three plots (and is the same as the spectral envelope of the $W_n C\#3$), but this envelope is filled with different set of harmonics in these two plots. Additional frequency lock-in is performed to make sure that the resulting sound will contain only the harmonics of the new pitch, making the sound clean and noise-free, by replacing phases of the coefficients of the Fourier transform of the sound wave in the restoration process by phases of the coefficients of the Fourier transform of the corresponding frequency pulse train. A few samples of music made with the new instrument are available on the web [12].

5. SUMMARY AND CONCLUSIONS

We developed and tested simple and powerful algorithms to perform separate modifications of pitch and timbre and to perform interpolation between sound samples. The algorithms is a new application of a cortical representation of the sound, which extracts perceptually important features similarly to the processing believed to occur in auditory pathways in primates, and thus can be used for making sound modifications tuned for and targeted to the ways the human nervous system processes information. We obtained promising results and are using algorithms in ongoing development of auditory user interfaces.

6. REFERENCES

[1] M. Elhilali, T. Chi, and S. Shamma (2002). "A spectro-temporal modulation index for assessment of speech intelligibility", Speech Communications, in press.

[2] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma (1999). "Spectro-temporal modulation transfer functions and speech intelligibility", J. Acoust. Soc. Am., vol. 106.

[3] M. Slaney, M. Covell and B. Lassiter (1996). "Automatic audio morphing", Proc. IEEE ICASSP 1996, Atlanta, GA.

[4] X. Serra (1997). "Musical sound modeling with sinusoids plus noise", in Musical Signal Processing, ed. by C. Roads et al., Swets & Zeitlinger Publishers, Lisse, The Netherlands.

[5] P. R. Cook (2002). "Real Sound Synthesis for Interactive Applications", A. K. Peters Ltd., Natick, MA.

[6] S. Barass (1996). "Sculpting a sound space with information properties: Organized sound", Cambridge University Press.

[7] D. Zotkin, R. Duraiswami, and L. Davis (2002). "Rendering localized spatial audio in a virtual auditory space", IEEE Trans. on Multimedia, in press.

[8] A. S. Bregman (1991). "Auditory scene analysis: The perceptual organization of sound", MIT Press, Cambridge, MA.

[9] N. Kowalski, D. Depireux, and S. Shamma (1996). "Analysis of dynamic spectra in ferret primary auditory cortex: Characteristics of single unit responses to moving ripple spectra", J. Neurophysiology, vol. 76(5).

[10] F. Jelinek (1998). "Statistical Methods for Speech Recognition", MIT Press, Cambridge, MA.

[11] http://www.isr.umd.edu/CAAR/

[12] http://dz.msk.ru/ICASSP2003/