Analysis of three pitch-shifting algorithms for different musical instruments

Anil Rai
Electrical Engineering Department
University of Bridgeport
Bridgeport, CT 06604, USA
anilrai@my.bridgeport.edu
*Corresponding author

Buket D Barkana
Electrical Engineering Department
University of Bridgeport
Bridgeport, Ct 06604, USA
bbarkana@bridgeport.edu

Abstract— **Pitch-shifting is a process where the original pitch of the sound is increased or decreased without affecting the length of the sound clip being recorded. Pitch shifters are being embedded in most of the audio processors. They are used to generate desired audio effects while recording sounds and music. Some of the notable uses of pitch shifters are in cartoons and entertainment industry for producing very distinct and unique voice, electronic music to generate melodies, metallic music to generate unnatural sound effects, and so on. However, one of the major problems of the pitch shifting algorithms is that there is not a sole algorithm, which can deal efficiently with all types of music and sounds. Here, we are evaluating the performance of three different pitch-shifting algorithms: pitch shifting via phase vocoder using phase locking, PitchshiftOcean, and pitch shifting via TSM using WSOLA for five musical instruments including guitar, piano, violin, drum, and flute. The objective of this work is to define the changes in some of the spectral and time-domain characteristics of musical instruments because of the pitch-shifting process.**

Keywords- pitch, phase vocoder, Short-time Fourier transform (STFT), Phase-Vocoder, analysis window, synthesis window, hop size, Time scale modification (TSM)

I. INTRODUCTION

Sound recording and music industry is one of the oldest and significant industries of today's age. It is also one of the most demanding industry in terms of effects required while recording. There are various kinds of effects in the sound recording techniques; time-based effects, modulation effects, pitch effects, filter effects, and so on. The audio effects are broadly classified into two types of effects as hardware and software effects. In our work, we will be dealing with a segment of software generated pitch shifting effects.

Pitch shifting is a process in an audio recording technique where the original pitch of the sound is increased or decreased without affecting the length of the sound clip being recorded. Pitch shifters are being embedded in most of the audio processors. They are used to generate desired audio effects while recording sounds and music. Pitch shifting is widely used in pitch correctors such as auto-tuning intonation errors while recording or performing the audios. Some of the most noted usages of pitch shifting are: generating various kinds of distinct and unique voices for animation characters, pitch correction, generating melodies in music, and generating unnatural audio effects in metallic music [1].

Pitch shifting can be done in both time-domain and frequency-domain. Some of the algorithms in pitching shifting in the time-domain are low latency audio pitch shifting, pitch shifting by Hilbert transform and IIR filters [2], pitch shifting by using TSM with various overlap-add methods [3] . Pitch shifting in the time-domain is characterized by low latency and low quality sounds. The major algorithms for pitch shifting in the frequency domain are pitch shifting using low latency pitch shifting in the frequency domain [4], phase vocoders [5], wavelets [6] and, pitch shifting using constant Q-transform [7]. The pitch shifting in the frequency domain is comparatively characterized by high latency and higher quality than the time domain. The main reason for this is the quality of modified sound is proportional to the use of DFT size, i.e. the higher the order of Fourier transform the better the quality.

Various kinds of musical instruments are played in the sound recording and music industry. Different instruments have various characteristics and they can be decomposed based on harmonicity, transient, sines, and percussive components. Moreover, harmonic features, tonality, and spectral flatness can be used to understand more about the nature of musical instruments. The music signal can be decomposed into the harmonic and percussive component, transient and non-transient components [6] [3] [8]. The major problem with pitch shifting algorithm is the absence of a single algorithm that performs efficiently on all kinds of sounds. Hence, in this paper we focus on the study of five different musical instruments and their changing spectral and time-domain characteristics with pitch shifting. The main goal of this work is to find the most suitable pitch shifting algorithms for different musical instruments, if possible.

The paper is organized as follows: Section II gives the details of the methodology. Section III presents the database. Section IV analyzes and discusses the experimental results. Section V concludes the paper.

II. METHODOLOGY

We analyzed seven different audio features in our work. They are calculated for various up and down pitch shifting

ratios to view the effects of pitch shifting algorithms in modified musical recordings of instruments. We compared the performances of three pitch-shifting algorithms: pitch shifting via phase vocoder using phase locking [9] [10], PitchshiftOcean [4], and pitch shifting via TSM using WSOLA [11] [3], by using several frequency- and time-domain features including tonality, spectral flatness, spectral centroid, spectral spread, harmonicity, harmonic component, percussive component, zero-crossing rate, and energy [12]. The outline of the methodology is depicted in Figure 1.

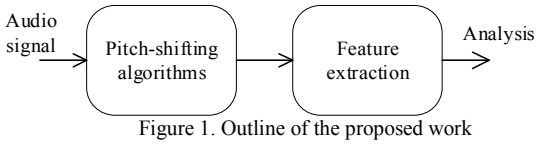

Figure 1. Outline of the proposed work

Pitch shifting is done by multiplication or division of every frequency component in a signal by a constant value called pitch shift ratio without affecting the duration of the signal. In frequency-domain, pitch shifting is performed by converting the time-domain audio signals into frequency-domain followed by scaling all the frequency bins by a pitch shift ratio while maintaining the phase of generated signals. It is usually done in two stages apart from some pre- and post- processing: analysis and synthesis stages. In the analysis stage, the time-domain signals are converted into frequency-domain in frame-to-frame format. These frames are pitch shifted. The pitch-shifted frames are reconstructed back into time-domain, which has the same length as the original signal, but its pitch is scaled.

There are various techniques available for pitch shifting in frequency domain. Techniques like phase vocoder and various improved versions of phase vocoder are used for pitch shifting. They are performed in two stages: the first stage resamples the signal and plays it in its original sampling frequency to produce pitch shifted and time stretched signal. The resulting time stretched, and pitch shifted signal can be time stretched again to recover the original length of the signal with the pitch shifted signal. This can be done by resampling signal with the rate of F(out)/F(in) and again stretching back with the time stretching factor of F(in)/F(out). The ratio of F(out)/F(in) is pitch shifting ratio [3]. The sounds modified this way have same artifacts as those generated by the time-scale modification (TSM) algorithm. They have also got some modified artificial sound called the chipmunk effect [13] [3]. Phase vocoder is one of the most commonly used time stretching techniques in frequency-domain. They employ the FFT to calculate the frequency bins. Improved phase vocoder is the modified version of phase vocoder solving phasiness by using STFT for time-frequency bins with phase-locking features in them. The purpose of the phase-locked vocoder is to reduce the vertical phase coherence using the fact each bin contributes partially. In the phase locking process, only frequency bins with spectral peaks are updated in the normal phase vocoder fashion and the remaining frequency bins are locked to the phase of the closest spectral peak.

A. Compared Pitch Shifting Algorithms:

We performed pitch-shifting between the range of 1-octave down and 1-octave up with an increment of 0.2 octaves. The shifting of one octave means the frequency component is doubled or halved without changing the tempo.

Method 1: Phase vocoder algorithm:

Phase vocoder employs the short-time Fourier transform (STFT) for estimating the instantaneous frequency.

$$X(m,k) = \sum_{r=-N/2}^{N/2} \mathrm{xm(r)w(\ r)\ exp(-2\pi ikr/N)} \qquad (1)$$

where m Є □ is the frame index, x_m is the m^{th} analysis frame, and w is the window function.

$$\mathrm{Tcoef(m)} = \frac{\mathrm{m.Ha}}{\mathrm{Fs}} \qquad (2)$$

Where *Ha* is the analysis hop size and *Fs* is sampling frequency in (2-3).

$$\mathrm{Fcoef(k)} = \frac{\mathrm{k.Fs}}{\mathrm{N}} \qquad (3)$$

It uses 2048-point FFT in its STFT and hop size of 512. The window type is Hanning window with a length of 2048.

Method 2: Pitch shift ocean algorithm:

It employs STFT for calculating the phasor Ω_x and that is followed by scaling. The Ω_a is copied into Ω_b, where Ω_b goes for scaling as in (3),

$$b = ka + 0.5 \qquad (4)$$

where *b* is modified frequency bin, *a* is current frequency bin, and *k* is the pitch shifting ratio. The modified phasor equation is given in (5),

$$\Omega_b = \Omega_b\, e^{\frac{2\pi(b-ma)p}{m.O.N}} \qquad (5)$$

where Ω_b is the modified phasor, *m* is the current frame number, *O* is the overlapping factor, and *N* is the DFT size.

It uses 2048-point FFT for analysis part and 4096-point FFT for synthesis part with hop size of 512 and 1024 each. The size of analysis and synthesis window are the same as FFT size. The window type is Hanning window.

Method 3: Pitch shifting using TSM WSOLA:

This is a time-domain approach of pitch shifting along with time stretching. In this approach, modified analysis frame x_m is used to generate synthesis frame y_m. Synthesis hop size is chosen in such a way that synthesis frames are overlapping and successive synthesis frames are aligned in a periodic format in an overlapping region [13] [14]. This method uses Hanning window with a size of 1024 and hop size of 512 along with 512 tolerance.

B. Analyzed features

Zero crossing rate (ZCR) is measure of the number of times the amplitude of the signal crosses the zero line. Periodic sounds have a smaller ZCR than their noisy counterparts. Average ZCR can give a coarse estimate of the spectral characteristics of the audio. Let s(*n*), where $n \in [0, N-1]$,

represent an audio frame of length N in the time-domain. ZCR can be calculated as in (6a-b).

$$ZCR_n = \sum_{m=0}^{N} |sgn[s(m)] - sgn[s(m-1)]| p(n-m) \quad (6a)$$

$$\begin{cases} p(n) = 1/2N & 0 \le n \le N-1 \\ 0 & otherwise \end{cases} \quad (6b)$$

The energy of audio signals is used to investigate the loss of frequency bins or an amplitude in modified audio signals.

Spectrum flatness is a numerical measure of the noisiness of the signal. It lies between (- ∞ dB, 0 dB) for each sound spectrum. Only a perfectly flat or constant spectrum can have a flatness of 0 dB. It can be used to check extra noisy components being added in a modified audio signal. The SFM measure is calculated by using the normalized log spectrum (7 and 8),

$$SFM = exp\left[\int_{-\pi}^{\pi} V(f) \frac{df}{2\pi}\right] \quad (7)$$

$$V(f) = \log\{\frac{|S(f)|^2}{E_x}\} \quad (8)$$

where *S(f)* is the discrete spectrum of *s(n)*. As the psychoacoustic parameter, tonality measures the perception of tonal content of the sound. Tonal sounds have a tonality measure close to 1. It is calculated as in (9). It is used to check the effect on the tonality of modified audio signals with respect to original signals.

$$Tonality = min\left(\frac{SFM_{db}}{-60}, 1\right) \quad (9)$$

Harmonic ratio (HR) is an estimate of harmonic components present in the spectrum. It is calculated as the maximum value of the autocorrelation of the signal frame [9]. It is used to check any deviations in harmonic component after audio modification.

Spectral centroid is considered as an estimate of the center of gravity of the spectrum within each subband. The sharpness of a sound is related to the spectral centroid that provides a noise-robust estimate of how the dominant frequency of the sound changes over time. It is measured as in (10). Spectral spread shows the spread of the spectrum around its mean value.

$$SCF_m = \frac{\sum_{f=l_m}^{u_m} f|W_m[f]S[f]|}{\sum_{f=l_m}^{u_m} |W_m[f]S[f]|} \quad (10)$$

The sound can be divided into M subbands, where frequency responses are $W_m(f)$, $m \in [1, M]$. l_m and u_m are the lowest and the highest frequencies in the m^{th} subband. The final SCF vector is obtained by concatenating all SCF_m values.

III. Database

Five instruments, guitar, piano, violin, drum, and flute, are tested. The time duration of each audio signal is 15 seconds with a sampling frequency of 44.1KHZ. Fig. 2 represents the audio signals in time-domain.

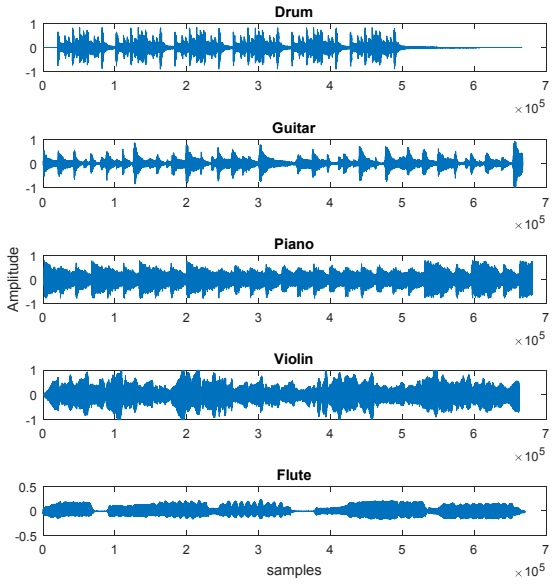

Figure 2. A 15-second audio of drum, guitar, piano, violin, and flute.

IV. Experimental Results, Analysis, And Discussion

The energy of the pitch shifted audio signals were measured to be close to the energy of the original audio signals. We did not observe any significant change in energy with different pitch shifting rates. We observed changes in zero-crossing rates, tonality, spectral spread, and spectral centroid measurements for all instruments. As we expected, ZCR increased when the pitch was shifted up and it decreased when the pitch was shifted down (Fig.3).

In terms of tonality, higher tonality was measured when the pitch was shifted down. Method 1 and 3 provided close tonality measures for all instruments with different pitch-shift ratios shown in Fig. 4. The highest gap between 1-octave down and 1-octave up shift was observed for the drum sound by using Method 1 and Method 3.

Except for the drum, all instruments' spectral flatness measure did not change much with up and down pitch-shifting. All three pitch shifting methods added small extra noisy component to the pitch shifted guitar, piano, violin, and flute audio. Method 3 added the most noise component to the drum sound with pitch shifting up. Method 2 caused the least amount of extra noisy component in drum sound Fig. 5.

Spectral spread and spectral centroid measures of the instruments increased when the pitch was shifted up. We observed a linear increase in these two measurements by using three methods, in which they provided similar measurements Fig. 6 and 7.

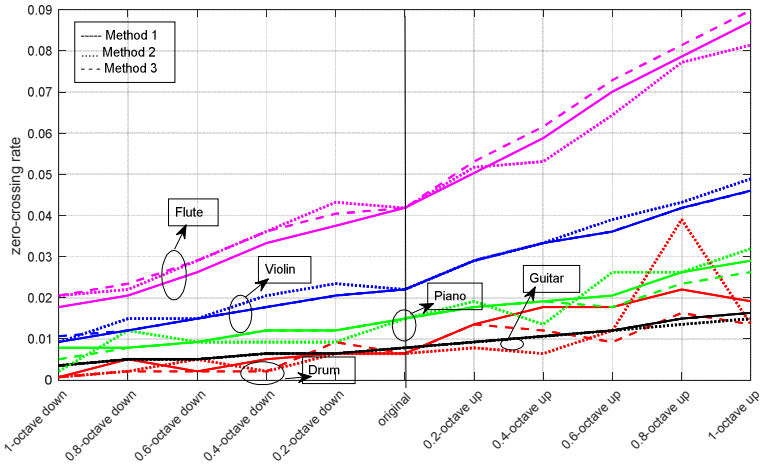

Figure 3. Zero-crossing rate (ZCR) measures of the guitar, piano, drum, violin, and flute sounds in the presence of pitch shifting.

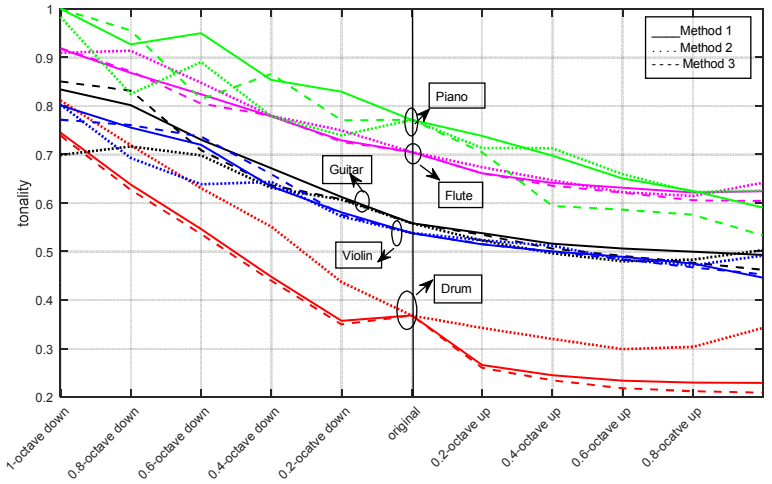

Figure 4.Tonality of the guitar, piano, drum, violin, and flute sounds in the presence of pitch shifting.

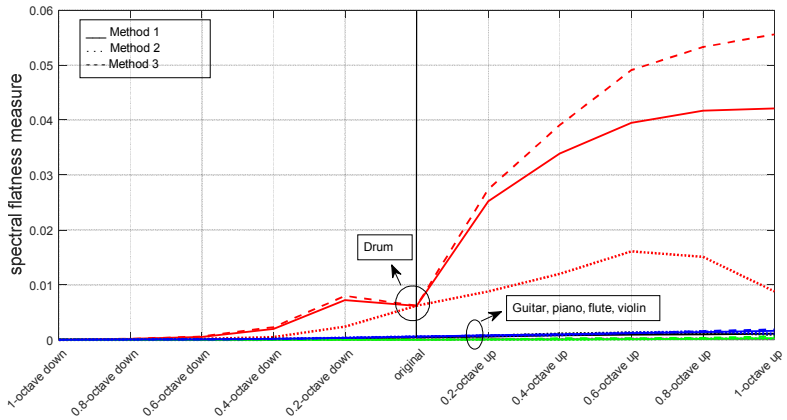

Figure 5.Spectral flatness measures (SFM) of the guitar, piano, drum, violin, and flute sounds in the presence of pitch shifting.

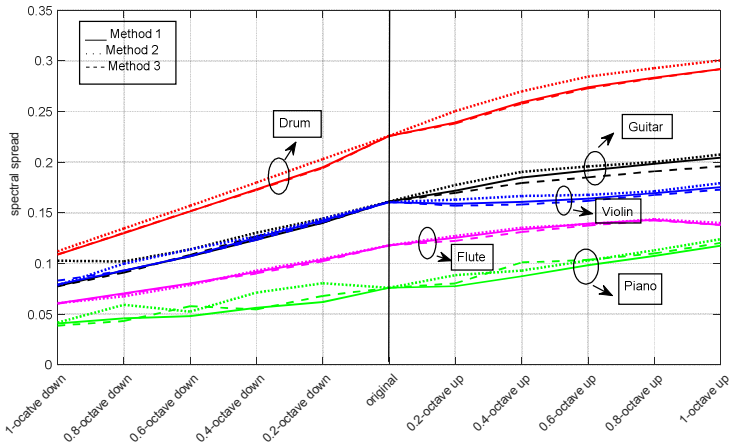

Figure 6. Spectral spread measures of the guitar, piano, drum, violin, and flute sounds in the presence of pitch shifting.

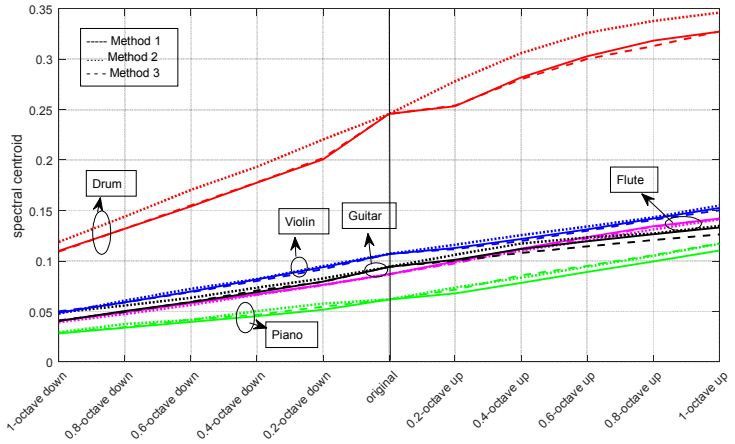

Figure 7. Spectral centroid measures of the guitar, piano, drum, violin, and flute sounds in the presence of pitch shifting.

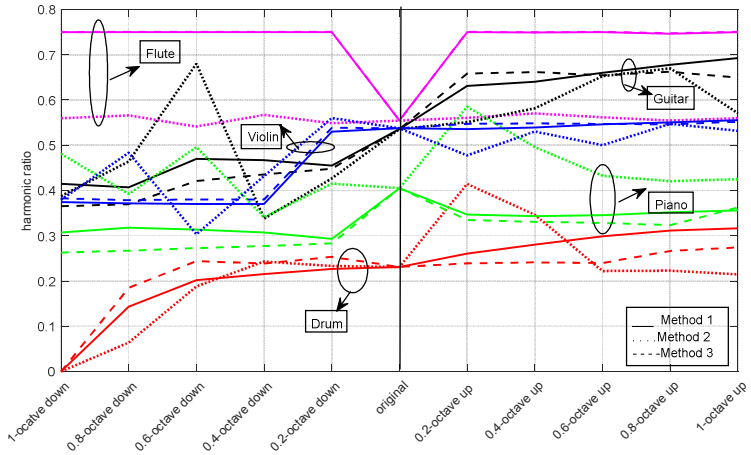

Figure 8. Harmonic ratio measures of the guitar, piano, drum, violin, and flute sounds in the presence of pitch shifting.

Harmonic ratios of the pitch-shifted instrument sounds were measured and a very slight change was observed between different pitch shifting ratios. Measurements by Method 1 and 3 were close to each other for all instruments. Harmonic ratios of the original audio of violin, drum, and piano were almost the same with the harmonic ratios of the 1-octave up versions of the audios by using three methods. Method 1 and 3 caused higher harmonic ratios for guitar and flute sounds in the presence of 1-octave up shifting while the Method 2 maintained the original harmonic ratio of the flute (Fig. 8).

V. Conclusion

This paper presented an analysis of three pitch-shifting algorithms: pitch shifting via phase vocoder using phase locking, PitchshiftOcean, and pitch shifting via TSM using WSOLA. Five musical instruments including guitar, piano, violin, drum, and flute were studied by using zero-crossing rate, spectral flatness, tonality, harmonic ratio, spectral centroid and spread, and energy. We observed the changes in these features after pitch shifting.

Our future works will investigate the similarities and differences between pitch-shifted sounds played by musicians and pitch-shifted sounds generated by pitch shifting algorithms.

References

[1] U. Zölzer, DAFX: Digital Audio Effects, John Wiley& Sons, 2nd Edition, 2011.

[2] S. Wardle, "A Hilbert transformer frequency shifter for audio," in First Workshop on Digital Audio Effects DAFx, 1998.

[3] J. Driedger, "Processing music signals using audio decomposition techniques", PhD Dissertation, May 2016.

[4] N. Juillerat and B. Hirsbrunner, "Low latency audio pitch shifting in the frequency domain," in 2010 International Conference on Audio, Language and Image Processing, pp. 16–24, 2010.

[5] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," IEEE Trans Speech Audio Process., vol. 7, pp. 323–332, 1999.

[6] A.G. Sklar, A Wavelet-based pitch-shifting method, http://umsis.miami.edu/~asklar/pitchshift.pdf, 2006.

[7] C. Schörkhuber, A. Klapuri, and A. Sontacchi, "Audio Pitch Shifting Using the Constant-Q Transform," J. Audio Eng. Soc., vol. 61, no. 7/8, pp. 562–572, Aug. 2013. [8] S. N. Levine and J. O. Smith III, "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications," presented at the Audio Engineering Society Convention 105, September 1998.

[9] P. N. Petkov and W. B. Kleijn, "Improving the Phase Vocoder Approach to Pitch-Shifting," INTERSPEECH, p. 4, 2007.

[10] T. Karrer, E. Lee, and J. O. Borchers, "PhaVoRIT: A Phase Vocoder for Real-Time Interactive Time-Stretching," in ICMC, 2006.

[11] D. Dorran, "Audio Time-Scale Modification," PhD Thesis, Dublin Institute of Technology, Jan. 2005.

[12] B. D. Barkana, N. John, and I. Saricicek, "Auditory Suspicious Event Databases: DASE and Bi-DASE," IEEE Access, vol. 6, pp. 33977–33985, 2018.

[13] J.-H. Chen, Audio time scale modification using decimation-based synchronized overlap-add algorithm," US7957960B2, 2011.

[14] W. Verhelst and M. Roelands, "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech," in proceedings of ICASSP-93, pp. 554–557, 1993.

Anil Rai (M'2019) is currently pursuing his M.S. degree in Electrical Engineering at University of Bridgeport. His research interests include audio signal processing, RF and microwaves.

Buket D. Barkana (M'2007) received the B.S. degree in electrical and electronics engineering from Anadolu University, Turkey, in 1994 and the M.S. and Ph.D. degrees in Eskisehir Osmangazi University, Turkey, in 2005. Her research interests include speech signal processing, environmental sound detection and classification system designs, and computer-aided diagnosis systems. She is currently a Professor of Electrical Engineering Department, University of Bridgeport, CT. She is the Director of the Signal Processing Research Group Laboratory with the Electrical Engineering Program, University of Bridgeport.