Presented at the Research Congress 2013
De La Salle University Manila
March 7-9, 2013

SPECTRAL FRAMING TECHNIQUE FOR AUDIO RECOGNITION AND CHARACTERIZATION

Roy Francis Navea
School of Engineering, De La Salle Canlubang

Abstract: This paper demonstrates a spectral framing technique for audio recognition and characterization. The spectral frequency and power density of the audio samples were studied to determine differences and similarities thus resulting to recognition and characterization.

Standard tones were used in order to test the algorithm for the spectral framing technique. The audio samples were taken from an electronic piano keyboard (musical instrument). Only a single octave (C_4 - C_5, where C_4 is the middle C) was considered in the study. The sounds or tones were recorded through the line-in terminal of the computer to minimize ambient noise and to obtain the full scale of the sample. A two-phase (training and testing) algorithm was made to characterize and test the audio samples. The Fast Fourier Transform (FFT) was used in order to convert the time-domain audio into its frequency domain. The study includes investigation of their power spectral densities which were divided into frames using the spectral sum and spectral hop (based on standard deviation) analysis. The spectral sum and hop were obtained by considering the number of frames close to the powers of 2 (2^n frames). The recognition and characterization criterions include spectral vector correlation, zero-difference, pitch dominant frequency and SPD frame pattern matching. The recognition for the framing technique is defined by identifying the minimum number of frames required to determine a correct pitch.

Results show that the methods applied were able to recognize the pitch of the tones using the above mentioned criterions.

Key Words: framing; vector coefficients; pitch; spectral power density, frame pattern

1. INTRODUCTION

Audio classification and segmentation can provide useful information for audio content understanding (Foote, 1997). A unique characteristic can be obtained by dividing an audio stream in segments which are composed of a number of elements. The concept of spectral framing goes in-line with segmentation. In this paper, the

segmentation varies. The divisions are defined by 2^n number of segments or frames.

Studies have been conducted on audio classification and segmentation and high-accuracy audio classification is achieved for simple cases such as speech/music discrimination (Lie et al, 2002). Most features used in this area are the same as the commonly used in speech and music processing which includes ZCR, MFCC, band energy, spectral centroid, BW, spectral flux, LPC and LPCC (Eronen et al, 2006, Tran et al 2009). The work of Karbasi et al (2011) even extended to environmental sound and extracts the spectral them to the frame-based classification using spectral dynamic features which processes changes throughout the frames of a sound file and appends spectral feature vectors as dynamic features.

In this paper, the main focus is finding an audio feature extraction algorithm for pitch detection using segmentation (framing). With an increasing number of frames defined by 2^n, this paper explored at the summation and standard deviation of the audio vector elements contained in each frame.

2. METHODOLOGY

2.1 Pre-Processing and System Training

The reference audio input was taken from a standard electronic piano keyboard. The designed system was programmed to record a 5-second audio sample and requires it to be a sustaining one. The sound of the accordion was chosen to produce a monotonic vibration. The study covered 13 tones only, C_4 (middle C) – C_5 as shown in Figure 1.

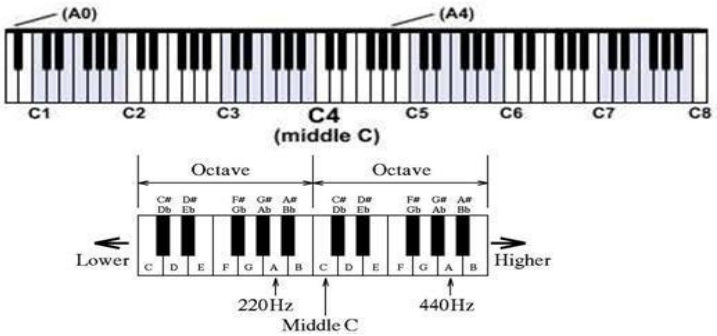

Figure 1. Standard Piano Keyboard. Retrieved January 25, 2013, from:
http://www.josef-k.net/mim/ThePianoKeyboard.gif &
http://0.tqn.com/d/piano/1/5/B/F/-/-/Scientific-Pitch-Notation_layout.png

The Fast Fourier Transform algorithm was used to show the frequency domain of

the audio input. For each tone, a set of FFT and spectral power density vector coefficients were drawn to serve as the basis for comparison later on. The unique dominant frequency for each tone was also derived.

2.2 FFT Vector Comparisons

The FFT vector coefficients obtained from the training process are stored in a database. This database is called every time an FFT vector comparison was made. Figure 2 shows the block diagram of the process. The audio sample was recorded and process using MATLAB® and the coefficients were exported to an external database.

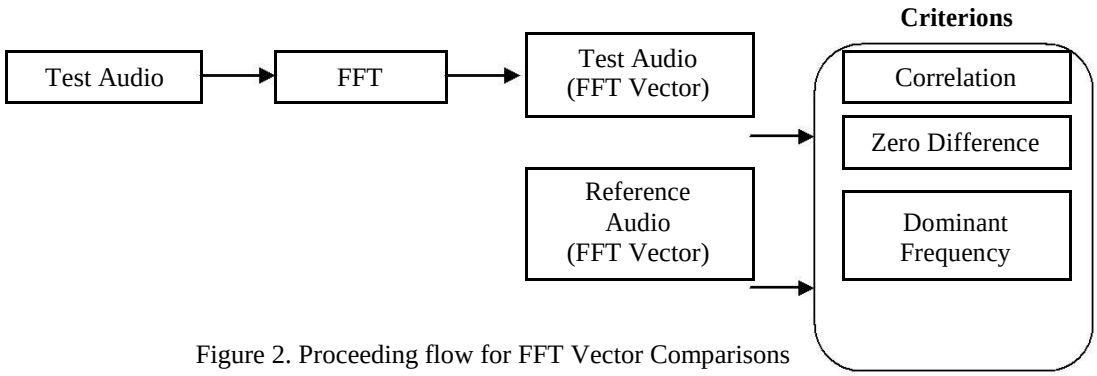

Figure 2. Proceeding flow for FFT Vector Comparisons

Three (3) criterions were used to identify the pitch of the test audio. The correlation coefficient is described as

(Eq. 1)

Where: r = correlation coefficient
x = reference audio vector
y = test audio vector
n = number of samples

The correlation coefficient between two signals measures the strength and the direction between two signals. The higher the correlation value, the closer the two signals are (Madain et al, 2010)

The zero-difference means that the two signals are closer if their difference and elemental vector sum is close to zero. In this study, the result of this algorithm is expressed in percentage and expressed as

——— **(Eq. 2)**

Where: %d = percentage difference from zero
x = reference audio vector
y = test audio vector
n = number of samples

Since the experimental setup for this study is closed within the given octave and secured by a line-in connection, the dominant frequency for each tone was considered. The database created also contains the dominant frequency of each tone. A simple matching algorithm was used to verify the pitch of the tone.

2.3 Framed SPD Vector Comparisons

The spectral framing technique is generally composed of two parts: SPD framing and feature extraction as in Karbasi et al (2011). There are a total of fourteen (14) frames used in this study. The number of frames was based on the integer-factor of the total number of samples immediately more than the n^{th} values of 2^n. Table 1 shows the frames and the number of samples for each frame.

Table 1. Frame Count and Samples per Frame

n	2^n	Number of Frames	Samples/Frame	n	2^n	Number of Frames	Samples/Frame
1	2	2	55125	**8**	256	294	375
2	4	5	22050	**9**	512	525	210
3	8	9	12250	**10**	1024	1050	105
4	16	18	6125	**11**	2048	2205	50
5	32	35	3150	**12**	4096	4410	25
6	64	70	1575	**13**	8192	11025	10
7	128	147	750	**14**	16384	18375	6

With reference to the work of Li (2000), the sum and standard deviation of each frame were calculated and the values obtained were used to form another vector. A frame pattern vector was used to contain the feature of the tone. It was formed by

taking the number of the n^{th} frame where the maximum occurrence of the sum and standard deviation occurred. This vector is described as

(Eq. 3)

Where: V_m = SPD frame pattern vector
S_i = frame where the maximum sum occurred
SD_i = frame where the maximum standard deviation occurred
i = n^{th} frame

All the feature vectors obtained from this section were subjected to the correlation and zero-difference criterions as in Figure 2, not including the dominant frequency matching.

3. RESULTS AND DISCUSSION

The recognition rate of the system using the FFT vector coefficients and dominant frequency matching is shown in Figure 3. The radar plot shows that the correlation and zero-difference criterions recognized the pitch of the tones all at 100%. On the other hand, the dominant frequency matching failed at some instances in identifying the pitch of the tone. Nevertheless, its recognition rate does not fall below 70%.

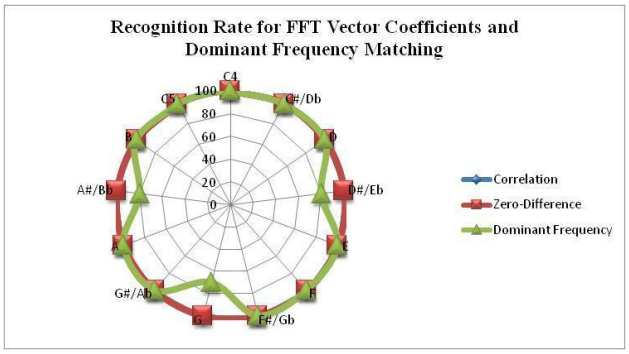

Figure 3. Recognition Rate for FFT Vector Coefficients and Dominant Frequency Matching

The results for the minimum number of frames required to identify the pitch of

the tone is shown in Figure 4. For the two methods used, the spectral hop required less number of frames as compared to the spectral sum. On the average, the spectral sum required 5.23 frames while the spectral hop was 4.23 frames. In the spectral sum framing, the top keys which were easily recognized are C4, F#/Gb, G, G#/Ab, A, B and C5. While in the spectral hop framing, the results show D, F, G, G#/Ab, and B.

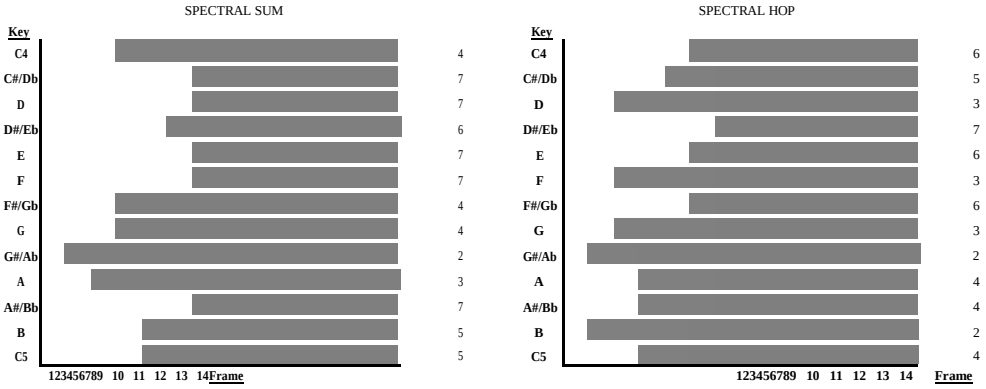

Figure 4. Minimum Frame Count for Pitch Detection

For the SPD frame pattern recognition, results show that the zero-difference criterion is better than the correlation criterion. The zero-difference criterion was able to identify all the pitch of the given tones while the correlation criterion showed low pitch recognition rates at some tones. Figure 5 shows that the zero-difference criterion completes the outer perimeter of the radar plot which shows 100% recognition. On the other hand, the correlation criterion fills some inner portion of the circle which shows that there are pitches that this criterion cannot identify at a higher percentage.

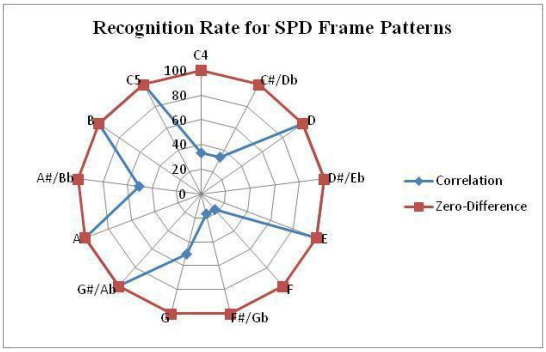

Figure 5. Recognition Rate for SPD Frame Patterns

4. CONCLUSIONS

This paper discussed the spectral framing technique for audio recognition and characterization. The technique was implemented for pitch recognition. Using the FFT vector coefficients of the audio sound, high recognition rates were achieved using the correlation and zero-difference criterions. The dominant frequency matching showed lower recognition rates. The spectral framing was implemented on the power density coefficients of the audio input. It was found out that the spectral hop required less number of frames as compared with the spectral sum before the pitch of the tone can be recognized. Last but the least, the zero-difference criterion results to higher recognition rates as compared to correlations when it comes to comparing the spectral frame patterns of the audio signals used.

Other several feature extraction techniques can be applied for pitch recognition and this could be furthered explored to cover the full scale of the musical tones.

5. REFERENCES

Eronen, A., Peltonen, V., Tuomi, J., Klapuri, S., Fagurlund, S., Sorsa, T., Lorho, G., & Huopaniemi, J. (2006). Audio-based context recognition, *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1

Foote, J. (1997). Content-based retrieval of music and audio, *proc. SPIE*, vol. 3229, pp. 138-147

Karbasi, M., Ahadi, S.M., & Bahmanian, M. (December 2011). Environmental sound classification using spectral dynamic features, *8th International Conference on Information, Communications and Signal Processing*, pp. 1-5

Li, S.Z. (September 2000). Content-based audio classification and retrieval using the nearest feature line method, *IEEE Transaction on Speech and Audio Processing*, vol. 8, pp. 619-625

Lie, L., Zhang, H.J. (2002). Content analysis for audio classification and segmentation, IEEE Transactions on Speech and Audio Processing vol. 10, no. 7

Madain, M., Al-Mosaiden, A., Al-Khassaweneh, M. (May 2010). Fault diagnosis in vehicle engines using sound recognition techniques, *IEEE International Conference on Electro/Information Technology*, pp. 1-4

Tran, H.D., & Li, H. (2009).Sound event classification based on feature integration, recursive feature elimination and structured classification, *proc. of IEEE ICASSP*