MULTI-TIMBRE CHORD CLASSIFICATION USING WAVELET TRANSFORM AND SELF-ORGANIZED MAP NEURAL NETWORKS

Borching Su and Shyh-Kang Jeng

Graduate Institute of Communication Engineering and

Department of Electrical Engineering

National Taiwan University

Taipei, Taiwan, ROC.

Email: skjeng@ew.ee.ntu.edu.tw

ABSTRACT

This paper presents a new method for musical chord recognition based on a model of human perception. We classify the chords directly from the sound without the information of timbres and notes. A wavelet-based transform as well as a self-organized map (SOM) neural network is adopted to imitate human ears and cerebra, respectively. The resultant system can classify chords very well even in a noisy environment.

1. INTRODUCTION

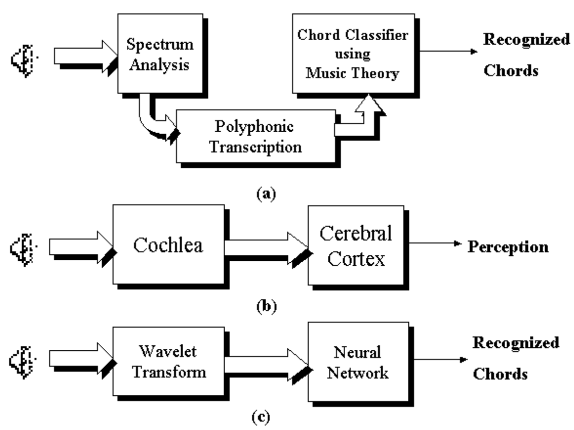


Fig.1 (a) Traditional chord recognition scheme. (b) Model of human perception to sounds. (c) Proposed system diagram.

Melodies, rhythms, and harmony are three fundamental components of music. For harmony in music the chords play an important role. Several chord recognition schemes have been developed by treating chords as the combination of discrete tones and recognizing them from the results of polyphonic analysis based on music theory [1]~[3]. A typical model of these scheme is shown in Fig.1(a). However, it does not fit our daily experience, since human beings often perceive chords as a whole with some readily recognized characteristics (e.g. major or minor) before they could accurately distinguish the individual notes composing the sound (Fig.1b). With this in mind, here we propose a model for direct chord identification in a multi-timbre environment (Fig.1c). The chord characteristics are extracted as a time-frequency map through a wavelet transform and then directly sent to a neural-network chord-classification unit without note identification. In next section, we will introduce some basic properties of musical timbres and chords. Implementation of the wavelet-transform and neural-network units will be introduced in Sections 3 and 4, respectively. Section 5 lists simulation results and gives related discussions. Finally, in Section 6 we draw some conclusions.

2. MUSICAL TIMBRES AND CHORDS

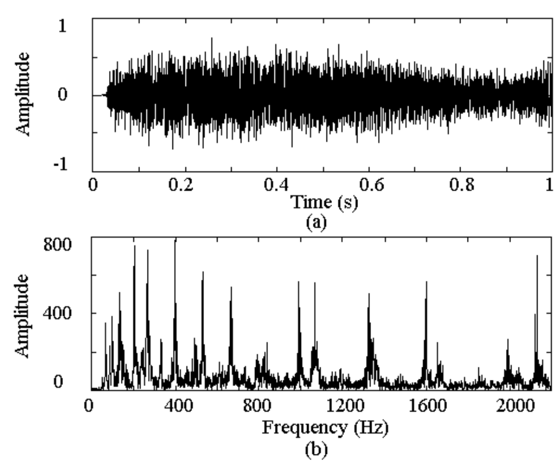


Fig.2 The first sound of the 4th movement of Beethoven's 5th Symphony. (a) Time domain signal. (b) Corresponding frequency spectrum.

	Frequency (Hz)	Equivalent MIDI No.	Closest MIDI Note
Fundamental Frequency	65.4064	24.0000	**C2**
1st partial	130.8128	36.0000	**C3**
2nd partial	196.2192	43.0196	G3
3rd partial	261.6256	48.0000	**C4**
4th partial	327.0320	51.8631	E4
5th partial	392.4383	55.0196	G4
6th partial	457.8447	57.6883	bB4
7th partial	523.2511	60.0000	**C5**
8th partial	588.6575	62.0391	D5
9th partial	654.0639	63.8631	E5

Table 1. A list of partials and equivalent MIDI numbers of C2.

Figure 2 (a) exhibits the first sound of the 4th movement of Beethoven's 5th symphony, consisting of 26 notes from 17 different kinds of instruments. It is hard for both human and machine to recognize all composing notes since various partials of various timbres overlap disorderly (Fig.2(b)). However, when a person listens to it, the sound in Fig.2 is with clear characteristic of a C major chord even though any of its composing notes is hard to detect.

Let's elaborate this point further. In frequency domain the partials for a specified timbre appear at frequencies approximately or equal to integer multiples of its fundamental frequency. Table 1 lists frequencies of the partials for note C2. Among these partials, some map exactly to octaves of the fundamental frequency, while others map to non-integer MIDI numbers. Here we let C4 = 262 Hz be the center C whose MIDI note number is 48. The closest MIDI notes of these partials are also listed. When a note of a timbre is played, all of its partials contribute to the time-frequency map and more or less hinder the recognition of notes.

As the number of notes and timbres increases, partials of all composing notes overlap disorderly. Most of them, especially those with a frequency/fundamental frequency ratio not equal to power of 2 will violate the rule of chords in music theory. This has been a serious problem in conventional polyphonic recognition [4][5][6].

3. WAVELET TRANSFORM

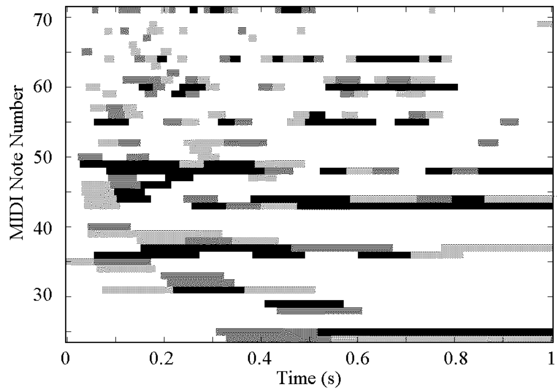


Fig.3 The time-frequency map of Figure 2.

This section shows the part of the system that simulates the role of human cochlea of human beings. Various schemes can be used for this goal, such as Short-time Fourier Transform (STFT), constant-Q filters, Wigner-Ville distribution, etc. [7]. Here we adopt the wavelet transform scheme since it has a "zooming" capacity over a logarithmic frequency range, and its translation-invariant property can center the sampling window properly in the time domain.

Several choices for the mother-wavelet $\psi(t)$ are available. In this research we apply a complex Gabor mother-wavelet, because it achieves the optimum of time and frequency localization [8, Chap.4]

$$\psi(t) = \exp\left(-\frac{t^2}{2} + j\omega_0 t\right) \qquad (1)$$

where ω_0 is the frequency of the mother-wavelet before it is scaled. In compliance with the musical requirement, we define the scaled versions of the mother-wavelet as

$$\psi_u^k(t) = \psi_{u,2^{k/v}}(t) = \frac{1}{\sqrt{2^{\frac{k}{v}}}}\psi\left(\frac{t-u}{2^{\frac{k}{v}}}\right) \qquad (2)$$

Here the index k represents the corresponding MIDI note number, u is the sampling time, and ν = 12 equals to the number of semitones in an octave. In order to relate k to MIDI notes, we set $\omega_0 = 2\pi *$ 16.352(Hz) for k=0, which is MIDI note C0 with a fundamental frequency 16.352(Hz). Using such wavelets, we can get the time-frequency map of Fig.2 as shown in Fig.3.

4. CLASSIFICATION AND TRAINING

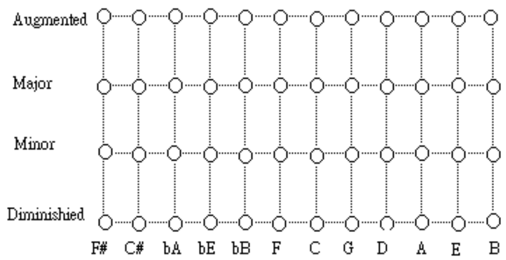


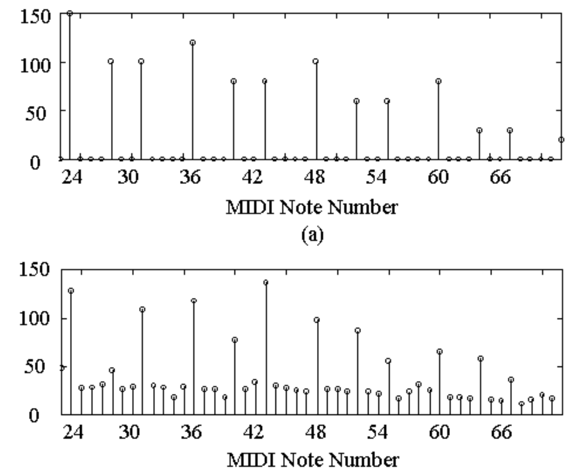
Fig.4 Self-Organized Map (SOM) for the chord classification. The horizontal axis refers to the tonality and the vertical axis represents the chords style.

As mentioned in Section 2 a chord is often with disordered partials such that the recognition of individual notes is very difficult. The neural networks can naturally lever this difficulty. Distinct chords present different characteristics in the time-frequency map, and the neural network can learn to classify them after training.

The neural network we adopt consists of a self-organized map layer. Two kinds of information should be determined to facilitate classification. One is the tonality, and the other is the chord style. These two kinds of information are chosen as the two dimensions of the self-organized map (SOM) shown in Fig. 4. In the tonality axis (horizontal), one of adjacent notes is dominant and the other is subdominant. In the chord style axis (vertical), adjacent styles are with two shared notes according to music theory. This configuration makes sure that adjacent neurons on the map are with high similarity.

Before learning, the initial synaptic weights of each neuron on the SOM are set according to music theory. Then a large number of training data extracted from real sounds are input to the network, and it starts to "experience" a chord. Since the SOM will learn from training data without any supervised information [9, Chap.9], the initial weights set above just give the map a pre-knowledge of the chords so that the network can converge more rapidly. Figure 5(a) shows a typical set of initial weights.

Three essential processes in training are competition, cooperation, and synaptic adaptation [9, Chap.9]. In the



(b)

Fig.5 Weights of the C major's neuron. The horizontal axis is MIDI note numbers. (a) Initial weights assigned according to music theory. (b) Final weights after training.

competition process, only one neuron among the 48 ones would be activated. In the cooperative process, the winning neuron tends to excite the neurons in its immediate neighborhood, which has a high similarity to the winning neuron. Finally, in the adaptive process, weights of neurons are gradually adjusted to fit the input patterns. Figure 5 (b) shows a typical trained set of weights.

5. RESULTS AND DISCUSSIONS

For training, 480 sound samples of 48 different kinds of chords have been used. The system then is ready for tested with recorded music segments. The recognition rate is defined as

$$\text{Recognition rate} = 1 - \frac{\text{number of incorrectly classfied}}{\text{total number of measures}}$$

The trained network is tested with the 4^{th} movement of Beethoven's 5^{th} Symphony conducted by Herbert von Karajan and performed by Berliner Philharmoniker in 1984. Fractional staff of the first 8 measures are shown in Fig.6 . According to music theory, chords of the eight measures are C major, Cmajor, Cmajor, Cmajor, Gmajor, Cmajor, Fmajor, Cmajor, respectively. The recognized chords fit all the 8 chords. Hence the recognition rate is 1 – 0 / 8 = 100%

Fig.6 The staff of Violins I and Basses of the first 8 measures of the 4th mov. of Beethoven's 5 th Symphony.

Amazingly, this recognition rate remains 100% even when we add a while Gaussian noise into the sound signal with a 0 dB signal-to-noise ratio (SNR). A recognition-rate to SNR plot as well as the 95% confidence intervals is shown in Fig.7.

This result shows the robustness of the system. Under a loud noise (SNR < -5dB), the recognition rate is kept at 75%, when individual notes are nearly unrecognizable. Since most trained humans can still tell such a sound as a faint impression of a chord, we may say this system has a "chord hearing" capability, which is similar to what a human being has.

6. CONCLUSIONS

We have developed a chord classification system using the wavelet transform as the "ear" and an SOM neural network as the "cerebrum." This system is extensible since chords not included can be easily added. With the capability of chord identification, we can do polyphonic recognition more accurately. This work can be an important building block in automatic transcription systems in the future. Results show that machine can directly "hear" the chords from a sound with a high recognition rate even under a noisy situation, as human beings do in a similar environment.

7. REFERENCES

[1] K.D. Martin, "A Blackboard System for Automatic Transcription of Simple Polyphonic Music," M.I.T. Media Lab Perceptual Computing Technical Report #385, July 1996.

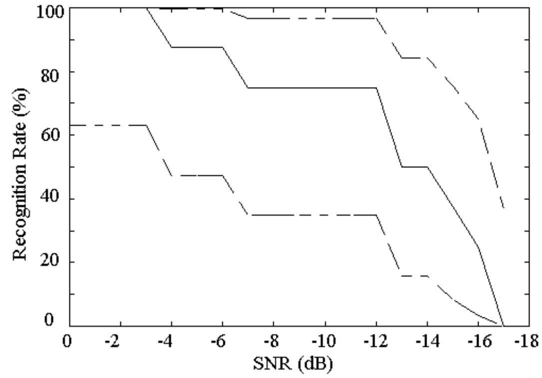


Fig.7 Recognition-rate vs. SNR plot. Dashed lines represent the 95% confidence intervals of corresponding recognition rates.

[2] K. Kashino, N. Hagita, "A music scene analysis system with the MRF-based information integration scheme," Proc. of the 13th International Conference on Pattern Recognition, vol.2, pp. 725 –729, 1996.

[3] H. Katayose, M. Imai, S. Inokuchi, "Sentiment extraction in music," 9th International Conference on Pattern Recognition, vol.2, pp. 1083 –1087, 1988.

[4] M.D. Macleod, "Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones," IEEE Trans. Signal Processing, vol.46, no.1, pp. 141 –148, Jan. 1998.

[5] P.J.Walmsley, S.J. Godsill, and P.J.W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1999.

[6] D.R. Franklin and J.F. Chicharo, "Paganini-a music analysis and recognition program," Proc. of the Fifth International Symposium on Signal Processing and Its Applications, vol. 1, pp. 107 –110, 1999.

[7] W.J. Pielemeier, G.H. Wakefield, and M.H. Simoni, "Time-frequency analysis of musical signals," Proc. IEEE, vol.84 no.9, pp.1216-1230, 1996.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press. 1999.

[9] S. Haykin, *Neural Networks - A Comprehensive Foundation*, Prentice Hall, 1999.