Azimuth Estimation in Polyphonic Music

Toni Heittola

Department of Signal Processing
Tampere University of Technology
P.O. Box 553, FI-33101 Tampere, Finland
e-mail: toni.heittola@tut.fi

1 Introduction

Most of the research in music information retrieval (MIR) is concentrating only on monophonic source signals. The stereo information has been ignored largely in the research many years. However, commercially available music recordings typically consist of a two-track stereo mix. The type of mixing process used in the recordings can be roughly categorizes into live recordings and studio recordings.

In live recordings, all musical instruments are usually recorded on a single stereo track using stereophonic microphone setup. The listeners localize sounds mainly based on time-differences between left and right channel, using the interaural time difference (ITD). In studio recordings, each musical instrument is recorded on a separate mono or stereo track. In the final mixing stage, audio effects (e.g. reverberation) can be added artificially. The virtual sound localization at any point between the left and right channel is achieved using proper amplitude for the left and right channel while mixing down tracks to a two-track stereo mix. Amplitude difference between channels is used to simulate interaural intensity difference (IID) by attenuating one channel and causing sound to be localized more in the opposite channel. The phase of a source is coherent between the channels. By assuming this mixing model, we can perform horizontal angle (azimuth) estimation for music signals [1, 2].

Azimuth information can be utilized in different applications of music information retrieval amongst musical instrument recognition and note streaming. In musical instrument recognition with polyphonic notes, the signal-to-noise ratio can be improved with beamforming in the feature extraction stage. Azimuth information can be utilized also in the note streaming of polyphonic audio, where notes can be grouped together based on pitch, timbre and azimuth.

Previously presented approaches to the azimuth estimations have been developed mainly for sound source separation based on predefined azimuth [1, 2]. We propose a system to estimate azimuths for individual notes in polyphonic music. Multipitch estimation is used to estimate fundamental frequencies (F0) for each note in the signal, and based on this F0 information azimuth estimation can be focused only on relevant harmonic components.

2 Azimuth Estimation

Most of the audio mixers use the sinusoidal energy-preserving panning law [4] to get the amplitudes of the left and right channels based on the panning value $\phi \epsilon \left[0, 1\right]$ as follows,

$$\begin{aligned} a_L &= \cos\left(\frac{\phi\pi}{2}\right), & (1) \\ a_R &= \sin\left(\frac{\phi\pi}{2}\right), & (2) \\ a_L^2 + a_R^2 &= 1 & (3) \end{aligned}$$

In the time-frequency domain, the channel signals $x_L(t)$ and $x_R(t)$ are denoted as $X_L(k,t)$ and $X_L(k,t)$, where k is the frequency index and t is the time index. By assuming previous panning law, azimuth can be estimated from the amplitude difference between left and right channel. The amplitude difference between channels is defined as

$$G(k,t) = \log\left(\frac{|X_R(k,t)|}{|X_L(k,t)|}\right) \tag{4}$$

By using the amplitude difference G, azimuth angle can be written as

$$azimuth\left(k,t\right) = \frac{360^\circ \cdot \arctan\left(\exp\left(G\left(k,t\right)\right)\right)}{\pi} - 90^\circ \tag{5}$$

In the proposed system, a time domain signal is divided into 100ms frames and the amplitude spectrum is extracted for each frame. Previously presented azimuth estimation is applied for each harmonic component of the note in question. Since overlapping notes may have overlapping frequency components as well, the harmonic components will have incoherent azimuth estimations. Overall azimuth estimation for the frame is achieved by taking the weighted median of estimated azimuths. Median is weighted with the amplitudes of the harmonic components to get more robust estimation.

Since azimuth can be assumed to be steady for individual notes, the median of the estimated azimuths for the frames is used to get the final estimation. Estimation is concentrated on high energy frames by weighting the median with the energy of harmonic components. Although overlapping harmonic components from other notes will destroy part of the azimuth information, the proposed approach still can capture enough information from the rest of the harmonic components within the note to give accurate azimuth estimation.

3 Experiments

Azimuth estimation was evaluated with simulated overlapping notes with random azimuths. Eight instrument classes (piano, electric piano, guitar, electric guitar, electric bass, saxophone, oboe, flute) were selected for the experiment. Thousand note combinations were randomly generated allowing all possible note and instrument combinations (excluding unison). Instrument samples were randomly selected from Real World Computing (RWC) database [3] and mixed together. Azimuths were randomly selected between -90° and 90° for each note. Multipitch estimation error was eliminated by using known fundamental frequencies of the individual notes in the estimation stage.

Azimuth estimation performance is presented in Table 1 for three different polyphonies, and for four different estimation accuracies (allowed absolute azimuth error). The proposed approach performed quite nicely with low note complexity, giving over 80 % accuracy with three note polyphony. Performance decreased with higher polyphony, giving still adequate estimation performance with lower estimation accuracy.

Table 1: Azimuth estimation results for multiple polyphony.

	Estimation accuracy			
Polyphony	$< 2.5°$	$< 5°$	$< 10°$	$< 20°$
3	81 %	84 %	86 %	89 %
6	59 %	64 %	69 %	75 %
9	47 %	53 %	60 %	67 %

References

[1] D. Barry, R. Lawlor, and E. Coyle. Sound source separation: Azimuth discrimination and resynthesis. In *Proc. of 7th International Conference on Digital Audio Effects*, pages 240–244, Naples, Italy, 2004.

[2] M. Cobos and J. J. López. Stereo audio source separation based on time–frequency masking and multilevel thresholding. *Digit. Signal Process.*, 18(6):960–976, 2008.

[3] M. Goto, T. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. pages 229–230, Baltimore, USA, October 2003. International Conference on Music Information Retrieval (ISMIR).

[4] M. Vinyes, J. Bonada, and A. Loscos. Demixing commercial music productions via human-assisted time-frequency masking. In *Proceedings of Audio Engineering Society 120th Convention*. Morgan Kaufmann, 2006.