IMPROVING SEPARATION OF HARMONIC SOURCES WITH ITERATIVE ESTIMATION OF SPATIAL CUES

Jinyu Han

Northwestern University
Evanston, IL 60208, USA
jinyuhan@gmail.com

Bryan Pardo[*]

Northwestern University
Evanston, IL 60208 USA
pardo@northwestern.edu

ABSTRACT

Recent work in source separation of two-channel mixtures has used spatial cues (cross-channel amplitude and phase difference coefficients) to estimate time-frequency masks for separating sources. As sources increasingly overlap in the time-frequency domain or the angle between sources decreases, these spatial cues become unreliable. We introduce a method to re-estimate the spatial cues for mixtures of harmonic sources. The newly estimated spatial cues are fed to the system to update each source estimate and the pitch estimate of each source. This iterative procedure is repeated until the difference between the current estimate of the spatial cues and the previous one is under a pre-set threshold. Results on a set of three-source mixtures of musical instruments show this approach significantly improves separation performance of two existing time-frequency masking systems.

Index Terms— Audio Source Separation, Harmonic Mask, Spatial Cues

1. INTRODUCTION

Audio source separation is the process of separating individual sources from mixtures of sources in an acoustic mixture. Effective audio source separation will lead to improvement in applications such as music remixing, sound identification and speech recognition. Here, we focus on the separation of harmonic sounds in a two-channel (stereo) anechoic mixture.

There are a number of different approaches to source separation, each depending on different assumptions. Suppose we have N sources and M mixtures. When M≥N, good source separation can often be achieved using Independent Component Analysis (ICA) [1]. When M<N, we have the underdetermined case, which makes traditional ICA ineffective. Stereo music (M=2) recordings tend to be underdetermined because they typically contain more than two sources (N>2).

Computational Auditory Scene Analysis (CASA) is inspired by Auditory Scene Analysis (ASA) [2], a perceptual theory that attempts to explain the remarkable capability of human auditory system. Several CASA systems have been developed for musical source separation [3][4], but their performance is limited. Recent CASA systems for monaural music separation achieved improved performance in situation where accurate pitch-tracking is possible [5][6][7].

Sparseness methods assume that there exists a representation of the sources such that the probability of two or more sources overlapping is low [8]. Provided this assumption, it is possible to blindly separate an arbitrary number of sources given just two anechoic mixtures. Richard and Yilmaz have proven that speech signals (in anechoic environments) are approximately disjoint in the time-frequency domain [9]. The DUET algorithm is a binary time-frequency masking technique for audio source separation that works well in this case [10].

The assumption that sources do not overlap significantly in the time-frequency domain does not always hold for music. Correlated onsets between sources and overlapping harmonics can render methods like DUET ineffective. The Active Source Estimation (ASE) algorithm was proposed to deal with harmonic sound sources that overlap in the time-frequency domain [11]. ASE builds on binary time-frequency masking approaches, such as DUET [10], by pitch tracking initial source estimates and assuming that energy should be present in the harmonics of each source (integer multiples of the fundamental frequency). Energy is redistributed among sources based on this assumption.

ASE and DUET are blind approaches that use spatial cues. While these methods are useful for audio source separation, their performances degrades as sources have more overlap or are placed close to each other (i.e. as the angle between sources decreases). Both cases result in inaccurate spatial cue estimates, making the sources difficult or impossible to separate.

In this paper, we introduce a method to iteratively improve source separation of time-frequency masking approaches that depend on spatial cues of harmonic sources. Section 2 provides a detailed description of the proposed method. Section 3 describes a dataset and experiment setup. Section 4 describes the results of applying our iterative improvement method.

2. SYSTEM DESCRIPTION

Our approach assumes an existing source separation approach is in use and sources in the mixture are harmonic. We begin with an overview of the approach. Given a mixture, we estimate the sources as follows:

1. Estimation of spatial cues for each source from the mixture signals. (Section 2.1)
2. Source estimation using a binary time-frequency masking algorithm (DUET or ASE) (Section 2.1).

[*] This work is funded by National Science Foundation Career Award (grant 0643752).

3. Pitch estimation on initial source estimates (Section 2.2).
4. Harmonic binary mask construction based on pitch information to identify the time-frequency bins where only one source has significant energy. (Section 2.3) (Section 2.4)
5. Energy extraction from the mixture using the harmonic binary masks so that each extracted signal only contains energy belonging to one source. (Section 2.4)
6. Spatial cues re-estimation on the signal from step 5 for each source. (Section 2.4)
7. Source re-estimation or output. (Section 2.5)
   - If the difference between the current and previous spatial cues estimates is under the threshold, output the source estimates.
   - Else, go to step 2.

We now provide some notational conventions. Let $X_1(\tau,\omega)$ and $X_2(\tau,\omega)$ be the time-frequency representation of two signal mixtures containing N source signals, $S_j(\tau,\omega)$, recorded by two omni-directional microphones.

$$X_1(\tau,\omega)=\sum_{j=1}^{N} S_j(\tau,\omega) \quad (1)$$

$$X_2(\tau,\omega)=\sum_{j=1}^{N} a_j e^{-i\omega\delta_j} S_j(\tau,\omega) \quad (2)$$

where a_j is the amplitude scaling coefficient and δ_j is the time-difference (inferred from phase differences) between the two microphones for the jth source, τ represents the center of a time window and ω represents a frequency of analysis of the time-frequency representation. We use $\hat{S}_j(\tau,\omega)$ to denote a source estimate.

2.1. Initial source estimation

Systems such as DUET and ASE, begin by grouping time-frequency frames based on spatial cues. Amplitude scaling $a(\tau,\omega)$ and phases difference $\delta(\tau,\omega)$ between the two mixtures in the stereo signal are calculated for every time-frequency frame. The intuition is that time-frequency frames with the same values for $a(\tau,\omega)$ and $\delta(\tau,\omega)$ are likely to come from the same source.

We first calculate the ratio $R(\tau,\omega)$ defined in (3) and then $a(\tau,\omega)$ and $\delta(\tau,\omega)$ in (4) and (5).

$$R(\tau,\omega)=\frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} \quad (3)$$

$$a(\tau,\omega)=|R(\tau,\omega)| \quad (4)$$

$$\delta(\tau,\omega)=\frac{-1}{\omega}\angle R(\tau,\omega) \quad (5)$$

where $/z/$ denotes the magnitude and $\angle z$ denotes the phase angle of a complex number.

The most common values for $a(\tau,\omega)$ and $\delta(\tau,\omega)$ can be found by creating a smoothed two dimensional histogram in the space of amplitude scaling and time-difference values. A K-means clustering algorithm is used to find the N most prominent peaks in the smoothed histogram. Each peak is assumed to correspond to one source in the mixture and the values for $a(\tau,\omega)$ and $\delta(\tau,\omega)$ at that peak are the mixing parameters for that source.

Once the mixing parameters for each source have been estimated, DUET assigns the energy in each time-frequency frame to the source whose peak lies closest to that frame in the space of a and δ. ASE works similarly, but uses a more sophisticated method to distribute energy in time-frequency frames that are distant from all peaks. In both cases, accurate estimation of the mixing parameters is vital to successful source separation. When the angle between sources is small or many time-frequency frames overlap, mixing parameter estimation becomes problematic.

2.2. Pitch estimation

Once we have created initial source estimates, we determine the fundamental frequency (pitch) of each source using an autocorrelation-based technique described in [12]. We denote the fundamental frequency of signal estimate $\hat{S}$ for time window τ as $\hat{F}(\tau)$. To smooth spurious, short-lived variation in $\hat{F}(\tau)$, any time-segment of less than 60ms that differs in pitch from the frames immediately before and after it by over 6% (roughly a semitone) is changed to match the pitch estimate in the previous frame.

2.3. Harmonic mask construction

ASE [11] builds on binary time-frequency approaches by pitch tracking initial source estimates and assuming energy should be present in the harmonics of each source. ASE then uses this knowledge to predict where harmonics of multiple sources may collide. We take a similar approach to predicting collisions.

We construct a harmonic mask $H(\tau,\omega)$ by finding each frequency ω at time τ that is (roughly) an integer multiple of the fundamental frequency estimate. We placing a 1 in each of these elements in the harmonic mask $H_j(\tau,\omega)$, as shown in (6).

$$H_j(\tau,\omega)=\begin{cases}1 & if \quad \exists k, |k\hat{F}_j(\tau)-\omega|<\Delta_g \\ 0 & else\end{cases} \quad (6)$$

Here, k is an integer and Δg is the maximal allowed difference in frequency from the kth harmonic of the source.

We call time-frequency frames with energy from multiple sources collision frames. We predict the location of collision frames by finding frames where there are multiple sources with a non-zero value in the harmonic mask.

The ASE system simply distributes energy in each collision frame among the source estimates based on the strength of each source's other harmonics in nearby time frames. This misses an opportunity to use non-collision frames for spatial cues re-estimation.

2.4. Spatial cues refinement

ASE (and DUET) work well only when the spatial cues estimate is accurate. However, when harmonic sources have too much overlap in the time-frequency domain or the angle

between instruments (as seen from the microphones) becomes small, these cues become unreliable. This is, in part, because the cues are calculated using an increasing number of collision frames. In this work, we take a similar approach to ASE in that we predict energy at the harmonics, but use these predictions to iteratively improve our spatial cues estimate.

Give n sources, we update $H_j(\tau,\omega)$ for source j using (7).

$$H_j'(\tau,\omega) = H_j(\tau,\omega) \times \prod_{i=1,i\neq j}^{n} (1 - H_i(\tau,\omega)) \quad (7)$$

This results in a mask that has value 1 in the time-frequency bins where we expect only source j to have energy. After we get the binary mask for each source, we lay mask $H_j'(\tau,\omega)$ on top of the mixture spectrums and extract the two-channel signals $M_j^1(\tau,\omega)$ and $M_j^2(\tau,\omega)$ which only contain energy from source j.

$$M_j^k(\tau,\omega) = X_k(\tau,\omega) \times H_j'(\tau,\omega) \quad (8)$$

where k is the channel (left or right) mixture for source j.

Although $M_j^k(\tau,\omega)$ may have poor fidelity to source j due to the missing energy from time-frequency frames containing overlap with other sources, the spatial cues estimated from this signal are more accurate than those estimated from the mixtures where sources interfere with each other.

We replace $X_1(\tau,\omega)$ and $X_2(\tau,\omega)$ in (3) with $M_j^1(\tau,\omega)$ and $M_j^2(\tau,\omega)$ respectively to estimate the spatial cues for source j. The refined spatial cues estimate for each source is then used to construct time-frequency masks which demix the mixtures, as described in section 2.1.

2.5. Final estimation

Every time the spatial cues estimate is updated, it is compared to the previous estimate to see if the difference between them is larger than a empirically set threshold. In our experiment, the threshold is set to be 5% of the previous spatial cues estimate, in both a and δ. If the difference in either dimension is larger than that, we feed the new spatial cues estimate back to the system, get new source and pitch estimates, and refine the spatial cues again. These procedures are repeated until the difference is smaller than the preset threshold. The source estimates in the last iteration are output as the final estimates. In our experiment, the iterative spatial cues estimation converges very quickly, usually taking three or four iterations.

3. EXPERIMENT

We evaluate performance on anechoic mixtures of three instruments. Each instrument plays a single note. The mixtures are major triads. We choose major triads because the harmonics show significant overlap in time and frequency and are an underdetermined case (more sources than mixtures).

The instrument recordings used in the testing mixtures are individual notes played by horn, bass clarinet and oboe, all taken from a set of instrument samples made available by the University of Iowa [13]. For each instrument, we have twelve pitches from C4 (roughly 262 Hz) through B4 (roughly 493 Hz), for a total of 12 recordings per instrument.

Mixtures of these recordings were created to simulate the stereo microphone pickup of spaced source sounds in an anechoic environment. We assume omni-directional microphones, spaced according to the highest frequency we expect to process, as in [10]. Instruments were placed in a semicircle around the microphone pair at a distance of one meter. The azimuth angle (mixing angle) between two adjacent instruments was varied from 15° to 50°. For each angle value, we created 30 mixtures, each of a block-chord major triad with one note per instrument. Notice that the closer two instruments are to each other, the more similar their spatial cues are and the more difficult to estimate the spatial cues and separate the sources from the mixture.

All sounds were normalized to have unit energy prior to mixing. Mixtures were created at 22.05 kHz and 16 bits, and were roughly one and half second in length. Time-frequency representations were created using a window length of 46ms and step size of 6ms for STFT processing.

4. EVALUATION

The estimated sources were compared to the original sources using the signal-to-noise ratio (SNR) shown in (9). SNR is widely used for blind audio source separation evaluation.

$$SNR = 10\log_{10}\left\{\frac{\sum_t s^2(t)}{\sum_t (\hat{s}(t) - s(t))^2}\right\} \quad (9)$$

where $s(t)$ and $\hat{s}(t)$ are the original and the estimated source signals, respectively.

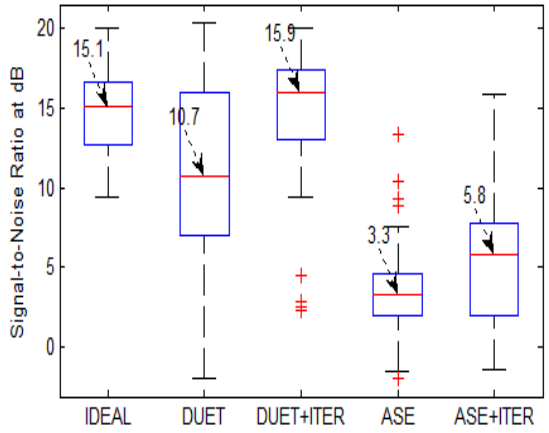

Figure 1: Performance of different methods at mixing angle 20°

Five systems are compared in our experiments. They are DUET with ground truth pitch information for spatial cues refinement (IDEAL), unmodified DUET, DUET with spatial cues iterative refinement (DUET+ITER), unmodified ASE and ASE with spatial cues iterative refinement (ASE+ITER). Ground truth pitches are the pitches estimated from the original recordings of the isolated sources, prior to mixing. Performance results when the angle between instruments is 20° are shown in

Figure 1. The median value of separation performance for each method is labeled with arrow text. Iterative spatial cues refinement improves DUET's median performance by 5.2 dB and ASE's median performance by 2.5 dB. Furthermore, our proposed system with iterative spatial cues estimation performs as well as the system using ground truth pitches.

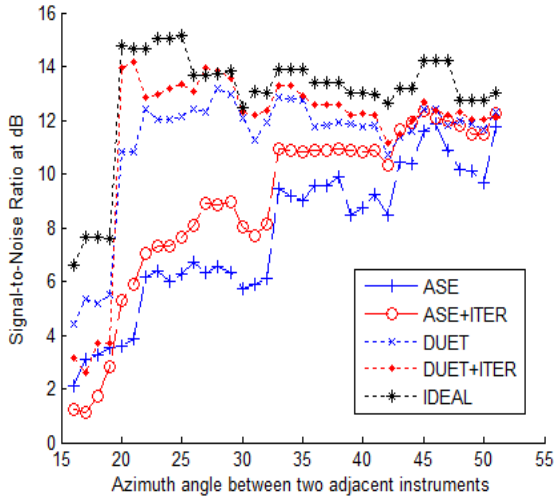

Figure 2: Signal to Noise Ratio of different methods with the mixing angle changing from 15° and 50°

Performance results from mixtures created using different mixing angle are shown in Figure 2. In this figure, each data point indicates an average result for 30 mixtures. The proposed system (DUET+ITER) consistently outperformed the existing systems' performance (excluding the system using ground truth pitches) for nearly all the mixing angles above 18°. When the mixing angle was below 18°, our proposed system performed poorly while the system using ground truth pitches achieved good results. This indicates that the sound sources are too close to each other, rendering pitch estimates from the initial source estimates inaccurate for spatial cues refinement. Otherwise, our iterative spatial cues estimation improves DUET or ASE when the sources are close to each other (In Figure 2, this is the case when the mixing angle is between 18° and 30°). As the angle increases, all of the systems' performance showed improving trends and our proposed system is nearly as good as the system using ground truth pitches when the angle is above 19°.

5. CONCLUSIONS

We have proposed a method for improved source separation of anechoic two-channel mixtures of harmonic sound sources. We use an existing source separation system to do the initial estimate and improve the results by incorporating the pitch and energy distribution information to further refine the spatial cues. Results on a database of three-instrument mixtures show this approach improves both the DUET and the ASE source separation systems, especially as the angle between two adjacent instruments falls below 40 degree.

6. REFERENCES

[1] J. V. Stone, *Independent Component Analysis: A Tutorial Introduction,* MIT Press, 2004.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of sound,* The MIT Press, 1990.

[3] D. K. Mellinger, *Event formation and separation in musical sound,* Ph.D. thesis, Stanford University, 1991.

[4] G. J. Brown and M. P. Cooke, "Perceptual grouping of musical sounds: A computational model," Journal of New Music Research, vol. 27, no. 4, pp. 107–132, 1999.

[5] Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, pp. 1475–1487, 2007.

[6] Y. Li and D. L. Wang, "Musical sound separation using pitch-based labeling and binary time-frequency masking," in *Proc. IEEE ICASSP*, 2008, pp.173-176.

[7] J. Woodruff, Y. Li, and D. L. Wang, "Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation," in *Proc. ISMIR,* 2008, pp. 538-543.

[8] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology,* vol. 15, pp. 18–33, 2005.

[9] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE ICASSP*, 2002, vol. 1, pp529-532.

[10] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, pp. 1830–1847, 2004.

[11] J. Woodruff and B. Pardo, "Using pitch, amplitude modulation and spatial cues for separation of harmonic instruments from stereo music recordings," *EURASIP Journal on Applied Signal Processing,* vol. 2007, 2007.

[12] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise of a sampled sound," In *Proc. the Institute of Phonetic Sciences,* 1993, vol. 17, pp. 97–110.

[13] http://theremin.music.uiowa.edu/index.html.