DEGENERATE UNMIXING ESTIMATION TECHNIQUE USING THE CONSTANT Q TRANSFORM

Zafar RAFII

Northwestern University
EECS Department
Evanston, IL, USA
zafarrafii@u.northwestern.edu

Bryan PARDO

Northwestern University
EECS Department
Evanston, IL, USA
pardo@northwestern.edu

ABSTRACT

The *Degenerate Unmixing Estimation Technique (DUET)* is a Blind Source Separation (BSS) algorithm for stereo audio. DUET depends on an amplitude-phase 2d histogram built from the differences between the two channels, where peaks in the histogram indicate sources in the mixture. If peaks overlap, separation becomes unfeasible. This is often the case for music mixtures. We propose to improve peak separation by building histograms from time-frequency representations based on the *Constant Q Transform (CQT)* instead of the Fourier Transform (FT). The CQT has a logarithmic frequency resolution matching the geometrically spaced notes of the Western music scale. We also adaptively resize histogram bins and use Wiener filtering to improve peak resolving and source reconstruction. Results on mixtures of harmonic musical instruments show improvement in separation, especially at low frequencies and for closely spaced sources.

Index Terms— Blind Source Separation, Degenerate Unmixing Estimation Technique, Constant Q Transform

1. INTRODUCTION

Blind Source Separation (BSS) is the separation of sources from mixtures without prior knowledge [1]. BSS finds uses in many audio-oriented tasks [1], such as speech/speaker recognition, vocalist/instrument identification, audio post-production, etc. For example, Independent Component Analysis (ICA) is a well-known family of BSS techniques which assumes that the sources are statistically independent [1]. However, ICA cannot be used when there are more sources than mixtures, a case referred to as "degenerate" [2].

Assuming that the sources can be represented sparsely in a given basis, sparse methods such as the *Degenerate Unmixing Estimation Technique (DUET)* can separate an arbitrary number of sources given a single stereo mixture [3]. DUET builds a 2d histogram from the ratio of the time-frequency representations between channels. Given a relatively anechoic mixture where time-frequency bins of different sources do not overlap too much, the histogram forms one peak for each source with peak location corresponding to the relative amplitude and phase parameters for that source. The mixture can then be partitioned by assigning each time-frequency bin to the source with the closest mixing parameters [2].

If there are too many time-frequency bins overlapping between different sources, the histogram cannot resolve peaks. This is often the case for music mixtures when using the Short-Time Fourier Transform (STFT). We improve peak separation by using a time-frequency representation based on the *Constant Q Transform (CQT)*. The CQT's logarithmic frequency resolution matches the geometrically spaced notes of the Western music scale [4]. This leads to fewer overlapping time-frequency bins between sources, especially in the lower octaves. We couple this with adaptive bin resizing for the histogram to further improve peak resolving and the use of Wiener filtering to improve source reconstruction.

Section 2 presents a review of the DUET and CQT methods. The contributions to the original DUET algorithm are presented in Section 3. Evaluation on mixtures of musical notes and harmonic instruments is conducted in Section 4. Finally, conclusion and perspectives are given in Section 5.

2. REVIEW

2.1. DUET

Given an anechoic stereo mixture recorded by two omnidirectional microphones, if a source k has a unique spatial location then it has a unique amplitude ratio α_k and phase difference δ_k between channels. Provided that the sources have sparse and disjoint time-frequency representations, the mixture can be partitioned by assigning each time-frequency bin to the source with the closest mixing parameters (α_k, δ_k).

To estimate the mixing parameters for each source, DUET computes for every time-frequency bin (τ, ω), the amplitude ratio $\alpha(\tau, \omega)$ and phase difference $\delta(\tau, \omega)$ between the STFTs of the stereo mixture. If many time-frequency bins share similar values of α and δ, they are likely to come from the same

source. The most common amplitude ratio and phase difference between channels are found by building a 2d histogram $H(\alpha, \delta)$ from $\alpha(\tau, \omega)$ and $\delta(\tau, \omega)$. Each peak in $H(\alpha, \delta)$ indicates a source with peak location corresponding to the estimated mixing parameters (α_k, δ_k) for that source. Binary time-frequency masks are then built to partition the STFTs of the mixture by assigning each time-frequency bin to the estimated (α_k, δ_k) which is closest to the local mixing parameters (α, δ) extracted for that bin. For more, see [5].

This method is particularly well suited to speech signals since their STFTs are sparse enough not to overlap too much when mixed, sufficiently for DUET to achieve good demixing results [6]. However, when too many time-frequency bins overlap between sources, peaks fuse in the 2d histogram so that peak/source separation becomes unfeasible. This is generally the case for music mixtures, simply because the STFT is not a time-frequency representation well adapted to music.

2.2. CQT

In modern Western music, the most common tuning system is the chromatic equal-tempered scale, which divides the octave into 12 logarithmically-spaced parts called semitones, with two adjacent semitones separated by a constant ratio of $2^{\frac{1}{12}}$ Hz. This is in line with the human auditory system which has a logarithmic frequency resolution [7]. Although highly efficient, the FT has frequency components separated by a constant difference. This forces a tradeoff: the Fourier Transform (FT) cannot get the needed frequency resolution at lower frequencies without significantly losing time resolution.

We need a transform with frequency components logarithmically spaced so that they match the notes of the twelve-tone equal-tempered scale. To be able to resolve adjacent notes played simultaneously, a quarter-tone spacing is needed (i.e. 24 frequency bins per octave). Unlike the FT, this transform should have a constant ratio Q of center frequency to resolution, leading to a logarithmic frequency resolution.

The *Constant Q Transform (CQT)* has these properties [4]. A fast implementation of the CQT exists, which makes use of the Fast Fourier Transform (FFT) in conjunction with a kernel, allowing the CQT to be as computationally efficient as the FFT [8]. Being an efficient transform more adapted to music mixtures, we therefore decided to use the CQT in combination with DUET. Note that, although the CQT has no inverse unlike the FT, it is needed only to build the 2d histogram. After estimation of the mixing parameters, synthesis is performed using the standard invertible STFT.

3. CONTRIBUTIONS

3.1. Contribution 1: Short-Time constant Q Transform

Similar to the way the STFT is built, we compute the Short-Time constant Q Transform (STQT) from the CQT of local segments using a sliding window of a fixed step size.

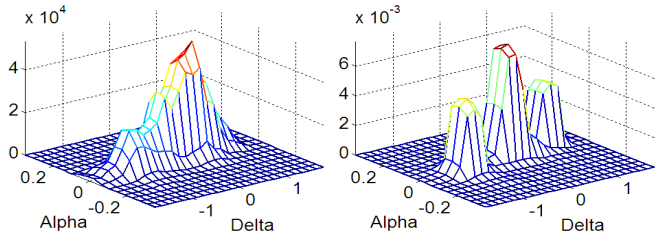

Fig. 1. 2d histograms of the mixture of the 3 piano notes $A2$, $B\flat2$ & $B2$, using DUET with STFT *(left)* and STQT *(right)*.

Figure 1 shows the 2d histograms of the mixture of the 3 piano notes $A2$, $B\flat2$ & $B2$, built using DUET with a STFT and default parameters detailed in [5] on the left, and DUET with STQT on the right. While the left histogram shows one gross peak because of the poor resolution of the FT at low octaves, the right histogram shows 3 clear peaks thanks to the log frequency resolution of the CQT, which can resolve peaks for adjacent pitches equally well in low and high octaves.

3.2. Contribution 2: Adaptive Boundaries

In [5], the 2d histogram is built using predefined boundaries and fixed-size bins. We propose the use of Adaptive Boundaries (AB) to automatically improve peak resolution when sources get too close to each other. To do so, we adjust the ranges of the α and δ values by analyzing their distributions and discarding outliers, on a case-by-case basis.

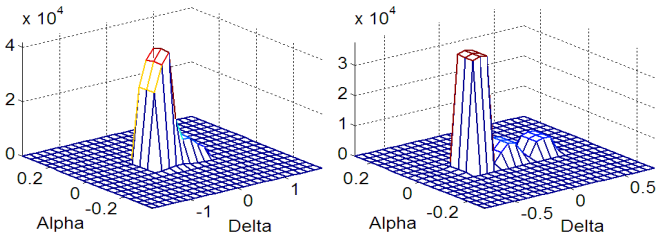

Fig. 2. 2d histograms of the mixture of the 3 piano notes $A6$, $B\flat6$ & $B6$, using DUET with STFT + fixed boundaries *(left)* and DUET with STFT + AB *(right)*

Figure 2 shows the 2d histograms of the mixture of the 3 piano notes $A6$, $B\flat6$ & $B6$, built using DUET with STFT + fixed boundaries on the left, and DUET with STFT + AB on the right. This time, the mixture has been synthesized by spacing the 3 sources closer to each other, so that the resolution needs to be refined. While the left histogram has one gross peak because of overly large boundaries/bins, the right histogram shows 3 clear peaks thanks to a finer resolution.

3.3. Contribution 3: Wiener Filtering

In [5], sources are reconstructed directly after partitioning the STFTs of the mixture, with the DC component discarded.

This will create masking artifacts, which can be mostly eliminated by adding back a little bit of the mixture to the demix, as suggested in [5]. However, that method creates interferences from unwanted sources in each estimated source. We propose to reconstruct the estimated sources by using a method based on a generalization of the Wiener Filtering (WF) for source estimation [9]. The method takes the magnitude spectrogram of the estimated sources and reconstruct the DC component, the symmetric part and the phase using the original STFTs of the mixture, giving a perceptually better and conservative separation, and reducing the interferences.

4. EVALUATION

4.1. Evaluation 1: Mixtures of Piano Notes

To evaluate the contributions proposed for DUET, we created two sets of mixtures of piano notes. The source recordings were from a SoundFont file provided by SONiVOX entitled "SB Stereo Piano V3.sf2". We used 85 half notes of 2 sec length from a grand piano sampled at 44,100 Hz, with pitches ranging from $A0$ (= 27.50 Hz) to $A7$ (= 3520 Hz).

Each mixture in the first set is the combination of 2 simultaneous notes. The 1^{st} pitch is always Ai and the 2^{nd} pitch is one of the 12 other higher pitches within the octave above it. The 12 intervals were generated for octave number i ranging from 0 to 6, for a total of 84 mixtures. Each mixture in the second set is the combination of 3 simultaneous notes. The 1^{st} pitch is always Ai, and the 2^{nd} and 3^{rd} pitches are such as the number of semitones between each pitch and the 1^{st} pitch is one of the following: (0,1,2), (0,2,4), (0,3,6), (0,4,8), (0,5,10), (0,6,12). Those 6 intervals were generated for octave number i ranging from 0 to 6, for a total of 42 mixtures.

Each combination of notes was mixed 5 times, once each with 5 different mixing angles between sources. The sources were placed on a circle of unit radius whose center corresponds to the location of two closely spaced microphones. The mixing angles were $(\pi - \frac{\pi}{12}j, \frac{\pi}{12}j)$ for the first set and $(\pi - \frac{\pi}{12}j, \frac{\pi}{2}, \frac{\pi}{12}j)$ for the second set, with j ranging from 1 to 5. This resulted in a total of 420 different mixtures of 2 piano notes and 210 different mixtures of 3 piano notes.

We evaluated the effectiveness of our contributions using DUET with default parameters [5] as our benchmark. The 2d histogram had 35 bins for α and 50 bins for δ, and weights $p = 1$ and $q = 0$, as suggested in [5]. Since there is no single appropriate technique for automatic peak location [5], we implemented a local maximum detector using a sliding window of size 3 by 3. We forced the peak enumeration to the highest local maxima, knowing a priori the number of sources.

To evaluate peak separability in the 2d histogram, we measured the mean Euclidean distance between ground truth peak locations of the original sources and corresponding estimates. To evaluate source reconstruction, we measured the mean Source to Distortion Ratio (SDR), Sources to Interferences Ratio (SIR) and Sources to Artifacts Ratio (SAR) between original sources and corresponding estimates [10].

As expected, results showed that the STQT improves peak and source separation, especially for octave numbers 0 to 2 ($\leq$ 200 Hz): Figure 3 shows the boxplots of the Euclidean distance as a function of the octave number. The AB improve peak resolving, so source separation, especially for small mixing angles ($\leq \frac{\pi}{6}$ rad): Figure 4 shows the SDR as a function of the angle between sources. The WF improves source reconstruction, reducing distortion and interferences (not shown here). A multivariate analysis of variance (MANOVA) showed that the results between the standard DUET and the enhanced DUET are statistically different.

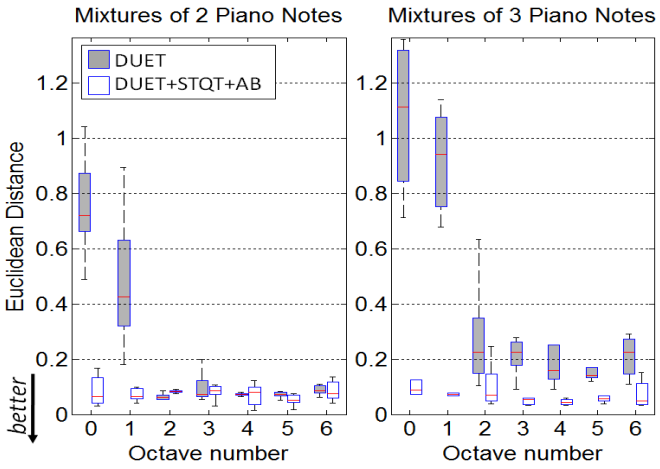

Fig. 3. Boxplots of the Euclidean distance between ground truth and estimated peak locations as a function of the octave number using the standard DUET and DUET with STQT + AB. Lower values are better. Outliers are not shown.

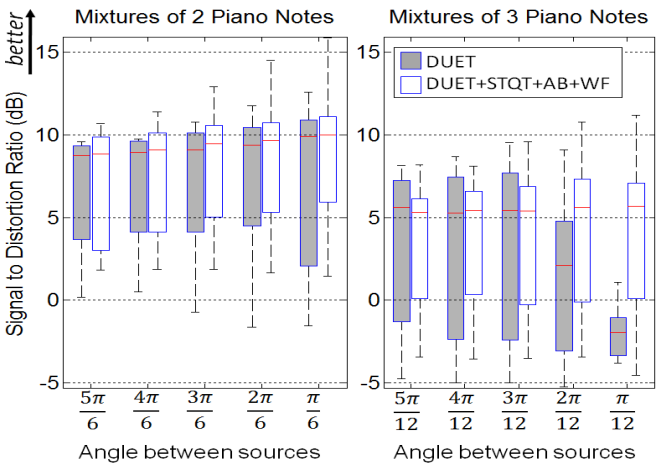

Fig. 4. Boxplots of the SDR as a function of the mixing angle using the standard DUET and DUET with STQT + AB + WF. Higher values are better. Outliers are not shown.

4.2. Evaluation 2: Mixtures of Harmonic Instruments

We then evaluated the contributions proposed for DUET on different sets of mixtures of harmonic instruments. We used 7 individual tracks of a classical recording of 14 sec length sampled at 44,100 Hz, downloaded from *ccMixter.org*, a community music site providing samples licensed under Creative Commons. The sources consist of synthesized instruments including soft strings, horns, bass, cello, violin and flute.

We used those 7 sources to create 5 sets of $\binom{7}{i}$ mixtures of all the possible combinations of i sources, with i ranging from 2 to 6, for a total of 119 mixtures of harmonic instruments. Each mixture was synthesized once assuming the i sources on a circle of unit radius whose center corresponds to the location of two closely spaced microphones, with the mixing angles being $\frac{j\pi}{i+1}$, with j ranging from 1 to i.

As in Section 4.1, we evaluated the different contributions using DUET with default parameters as our benchmark. This time, the 2d histogram had 70 bins for α and 100 for δ, and weights $p = 0.5$ and $q = 0$ [5]. We used a local maximum detector with peak enumeration forced to the i highest local maxima. We measured peak separability using the Euclidean distance and source reconstruction using SDR, SIR and SAR.

Results confirmed that DUET with STQT + AB + WF improves peak separation, peak resolving and source reconstruction, for up to 6 sources: Figure 5 shows the Euclidean distance and the SDR as a function of the number of sources. MANOVA showed that the results between the standard DUET and the enhanced DUET are statistically different.

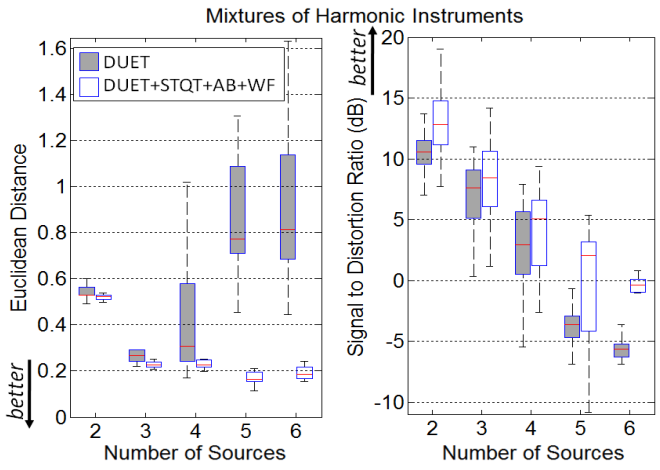

Fig. 5. Boxplots of the Euclidean distance and SDR as a function of the number of sources using the standard DUET and DUET with STQT + AB + WF. Outliers are not shown.

5. CONCLUSION

We proposed contributions to improve DUET. Experiments on mixtures of piano notes showed that time-frequency representations based on the CQT improve peak/source separation, especially up to low frequencies ($\leq$ 200 Hz), adaptive boundaries improves peak resolving, especially when sources are closely spaced ($\leq \frac{\pi}{6}$ rad), and Wiener filtering improves source reconstruction. Experiments on mixtures of harmonic instruments confirmed those improvements, up to 6 sources. Additional experiments showed that CQT gives equally well results on mixtures of female and male speech.

This work was supported by NSF grant numbers IIS-0643752 and IIS-0757544.

6. REFERENCES

[1] Pierre Common and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[2] Alexander Jourjine, Scott Rickard, and Özgür Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2985–2988, Istanbul, Turkey, June 2000.

[3] Paul D. O'Grady, Barak A. Pearlmutter, and Scott T. Rickard. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, 15(1):18–33, July 2005.

[4] Judith C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, January 1991.

[5] Scott Rickard. *The DUET Blind Source Separation Algorithm*, pages 217–241. Springer Netherlands, 2007.

[6] Özgür Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. In *IEEE Transactions on Signal Processing*, volume 52, pages 1830–1847, July 2004.

[7] Diana Deutsch. *The Psychology of Music, second edition*. Academic Press, San Diego, 1999.

[8] Judith C. Brown and Miller S. Puckette. An efficient algorithm for the calculation of a constant Q transform. *The Journal of the Acoustical Society of America*, 92(5):2698–2701, November 1992.

[9] Laurent Benaroya, Frédéric Bimbot, and Rémi Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, Language Processing*, 14(1):191–199, January 2006.

[10] Cedric Févotte, Rémi Gribonval, and Emmanuel Vincent. BSS EVAL toolbox user guide. Technical Report 1706, IRISA, Rennes, France, April 2005. http://www.irisa.fr/metiss/bss eval/.