36th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011) May 22-27, 2011, Prague, Czech Republic

NORTHWESTERN UNIVERSITY

Degenerate Unmixing Estimation Technique using the Constant Q Transform

Zafar RAFII & Bryan PARDO

Northwestern University, EECS Department, Evanston, IL, USA.

zafarrafii@u.northwestern.edu • *pardo@northwestern.edu* • *http://music.cs.northwestern.edu/*

interactive audio lab

Introduction

DUET is a blind source separation algorithm which depends on an amplitude-phase 2d histogram built from the differences between the two channels of a stereo mixture, where peaks in the histogram indicate sources in the mixture. With music mixtures, peaks often overlap, which makes the separation unfeasible. We propose to build histograms from time-frequency representations based on the CQT to improve peak/source separation. We also use adaptive boundaries and Wiener filtering to improve peak resolving and source reconstruction.

Review

▸ Degenerate Unmixing Estimation Technique (DUET)

DUET depends on building a 2d histogram from the ratio between channels, where peaks represent the relative amplitude α and phase δ parameters for each source. However, sources in music mixtures often overlap in time-frequency when using the Short-Time Fourier Transform (STFT).

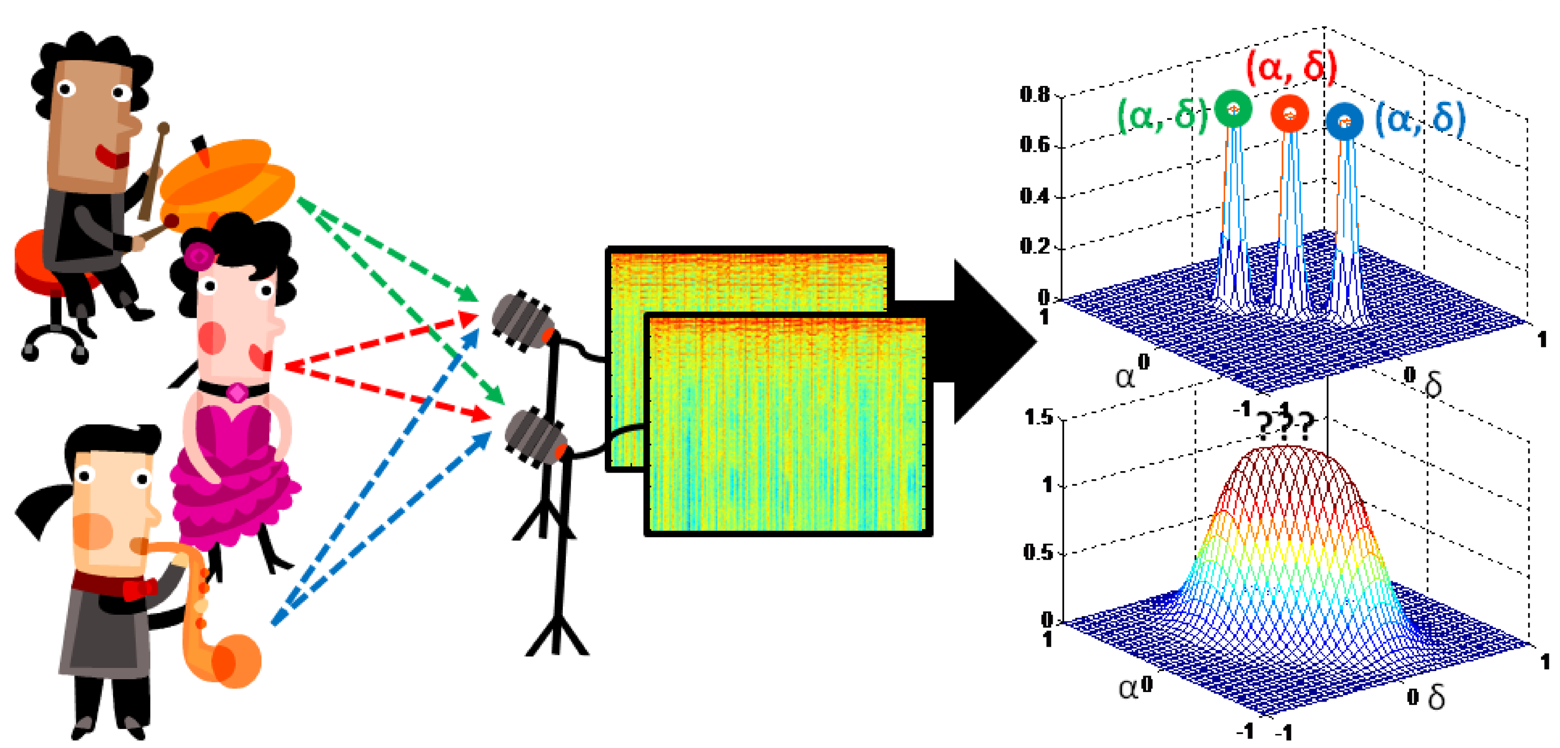

Figure: DUET builds a 2d histogram from the ratio of the STFTs between channels. The top histogram is what we should get (one peak for each source), but the bottom histogram is what we often get with music mixtures (one unworkable large peak).

▸ Constant Q Transform (CQT)

Unlike the Fourier Transform (FT), the CQT has a logarithmic frequency resolution matching the geometrically spaced notes of the Western music scale, therefore better adapted to music mixtures.

Contributions

▸ Short-Time constant Q Transform (STQT)

Instead of using the STFT to build the 2d histogram, we use the STQT, similarly built from the CQT of local segments using a sliding window.

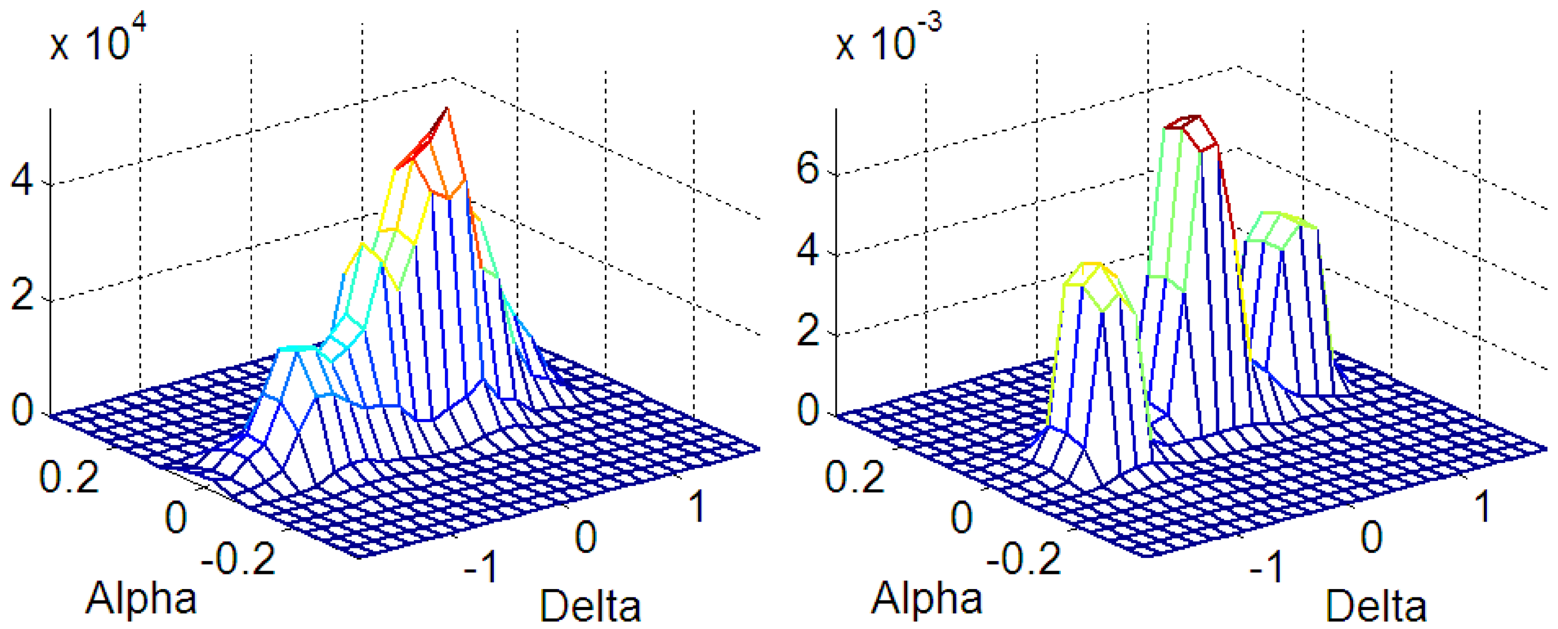

Figure: 2d histograms of the mixture of the 3 piano notes $A2$, $B\flat2$ & $B2$, using DUET with STFT *(left)* and DUET with STQT *(right)*.

▸ Adaptive Boundaries (AB)

Instead of using fixed boundaries for the 2d histogram, we automatically adapt the ranges of values for α and δ based on their distributions.

▸ Wiener Filtering (WF)

Instead of reconstructing the sources directly after masking, we use Wiener filtering to reduce distortion, interferences and masking artifacts.

Evaluation

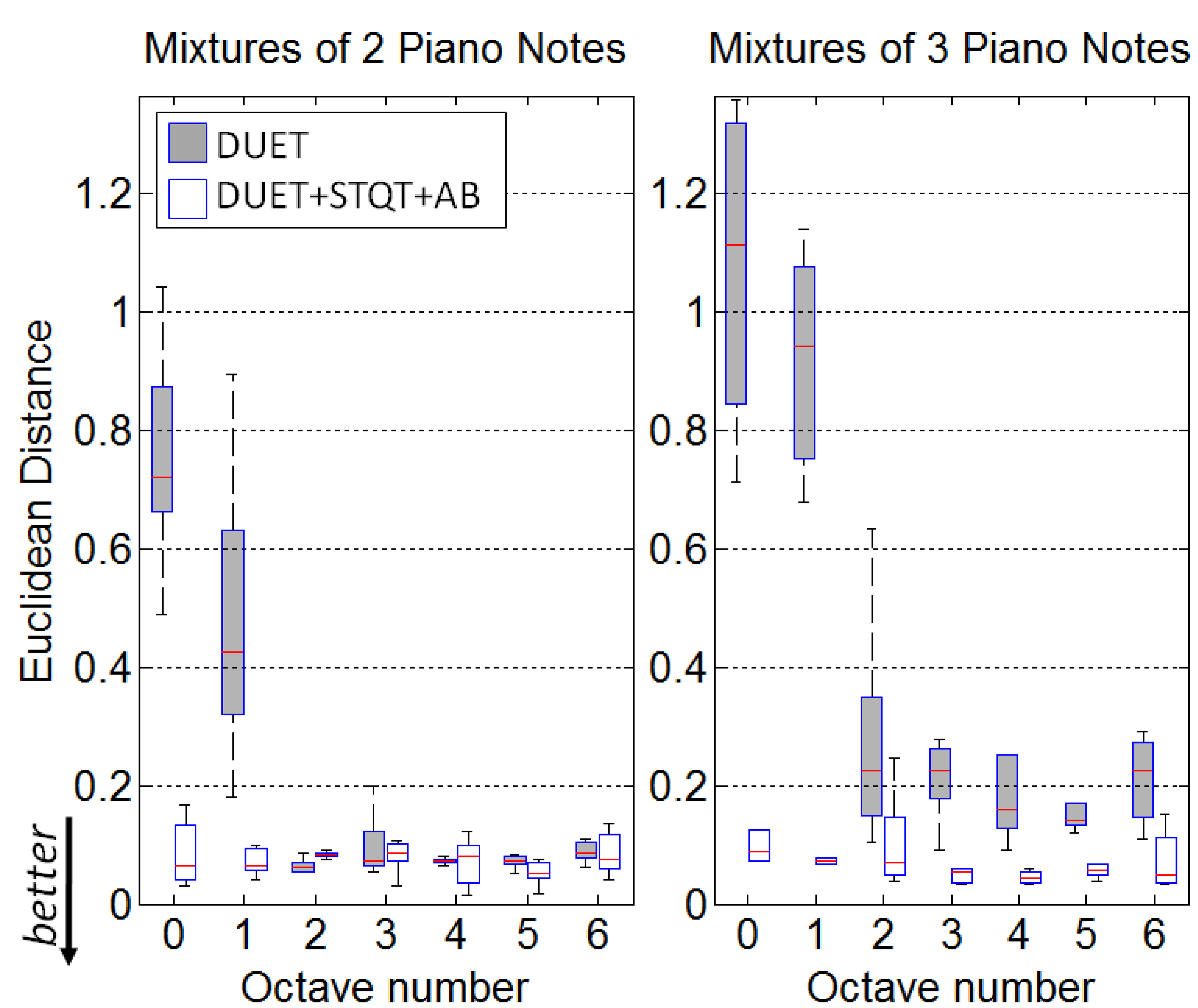

Figure: Box plots of the Euclidean distance between ground truth and estimated peak locations as a function of the octave number using DUET and DUET+STQT+AB. Lower values are better. Outliers are not shown.

▸ Mixtures of Piano Notes

- 85 half notes of 2 sec length from a grand piano
- pitches ranging from $A0$ (= 27.50 Hz) to $A7$ (= 3,520 Hz)
- 84 mixtures of 2 notes at 7 different octave ranges
- 42 mixtures of 3 notes at 7 different octave ranges
- 5 different mixing angles

▸ Mixtures of Harmonic Instruments

- 7 individual tracks of 14 sec length from a classical recording
- instruments including cellos, flute, horns, violas, and violins
- 5 sets of mixtures, from 2 to 6 sources, for a total of 119 mixtures

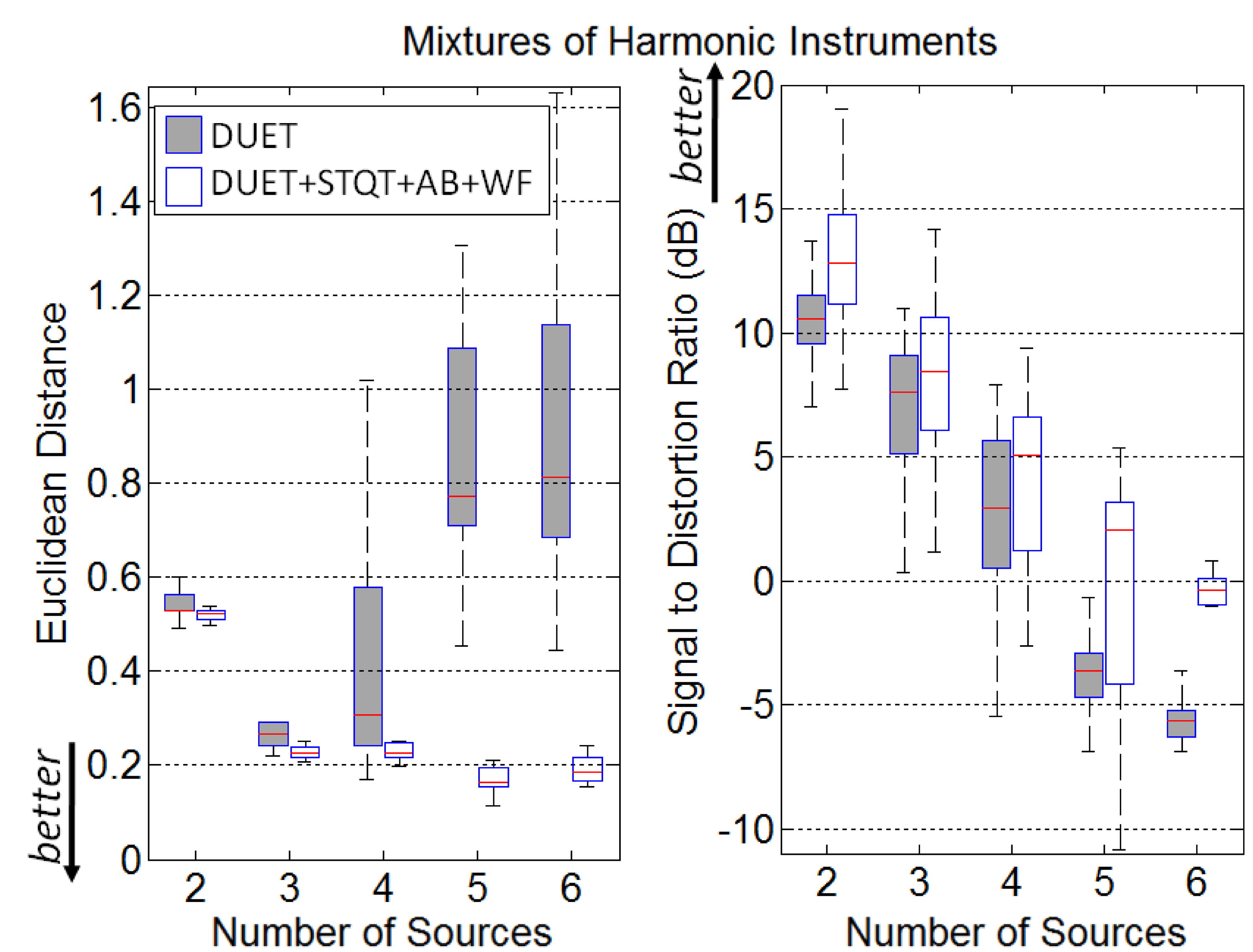

Figure: Box plots of the Euclidean distance between ground truth and estimated peak locations and Signal-to-Distortion Ratio (SDR) as a function of the number of sources using DUET and DUET+STQT+AB+WF. Outliers are not shown.

Conclusion

Experiments with mixtures of piano notes showed that (1) CQT improves peak separation especially for low frequencies ($\leq$ 200 Hz), (2) adaptive boundaries improve peak resolving especially for close sources ($\leq \frac{\pi}{6}$ rad), and (3) Wiener filtering improves source reconstruction. Experiments with mixtures of harmonic instruments confirmed those improvements. Additional experiments showed that CQT gives equally well results with mixtures of female and male speech.

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