A Novel Cepstral Representation for Timbre Modeling of Sound Sources in Polyphonic Mixtures

Zhiyao Duan[1], *Bryan Pardo*[2], *Laurent Daudet*[3]

[1]University of Rochester, NY, USA. • [2]Northwestern University, IL, USA. • [3]Université Paris Diderot - Paris 7, Paris, France.
zhiyao.duan@rochester.edu • *pardo@northwestern.edu* • *laurent.daudet@espci.fr*

Introduction

▶ Proposes a new cepstral representation named Uniform Discrete Cepstrum (UDC) and its mel-scale variant (MUDC).

Ordinary Cepstrum (OC)	- calculated from the full
MFCC	magnitude spectrum
Discrete Cepstrum (DC)	- calculated from isolated
Regularized DC (RDC)	spectral points
Uniform DC (UDC)	- can model timbre of sound
Mel-frequency UDC (MUDC)	sources w/o source separation

▶ Derives mathematical relations between these representations.

Relations of Cepstral Representations

▶ Cepstrum: approximate log-amp spectrum with sinusoids

$$a(f) \approx c_0 + \sqrt{2}\sum_{i=1}^{p-1} c_i \cos(2\pi i f) \quad (1)$$

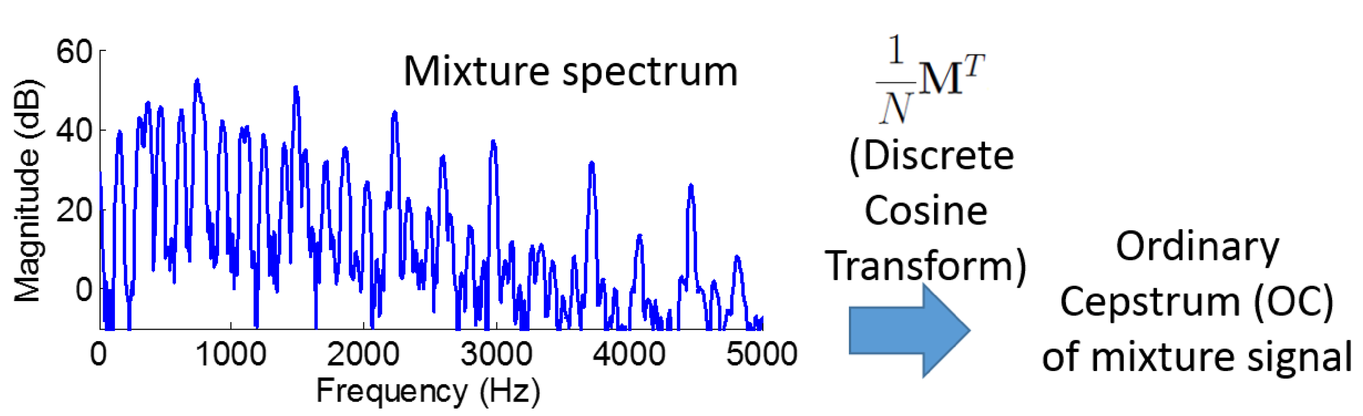
- $a(f)$: log-amplitude spectrum at normalized frequency f
- c_i's : cepstral coefficients of order p

▶ Ordinary Cepstrum (OC): least square solution of Eq.(1)

$$\mathbf{a} = \mathbf{M}\mathbf{c} \quad (2)$$

$$\mathbf{c}_{\text{oc}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{a} = \frac{1}{N}\mathbf{M}^T\mathbf{a} \quad (3)$$

- $\mathbf{a}$: discrete log-amplitude spectrum with N frequency bins
- $\mathbf{M}$: the first p columns of a discrete cosine transform (DCT) matrix whose columns are orthogonal, so $\mathbf{M}^T\mathbf{M} = \frac{1}{N}\mathbf{I}$.



▶ MFCC: apply mel-scale filterbank before DCT

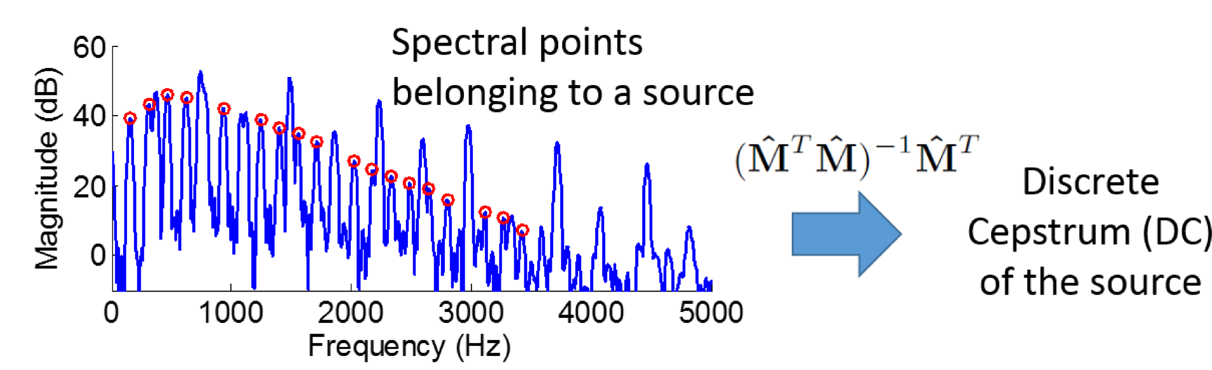
- Still approximating the mixture spectrum

▶ DC: least square solution approximating selected points

$$\hat{\mathbf{a}} = \hat{\mathbf{M}}\mathbf{c} \quad (4)$$

$$\mathbf{c}_{\text{dc}} = (\hat{\mathbf{M}}^T\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}^T\hat{\mathbf{a}} \quad (5)$$

- $\hat{\mathbf{a}}$: log-amplitudes of selected spectral points likely belonging to a source (e.g., harmonics of a pitched source)
- $\hat{\mathbf{M}}$: rows of $\mathbf{M}$ corresponding to the spectral points
- $(\hat{\mathbf{M}}^T\hat{\mathbf{M}})^{-1}$ is often poorly-conditioned due to large frequency gap between these points, i.e., approximated spectral envelope overfits these points and oscillates significantly at other frequencies.



▶ RDC: adding a regularizer to prevent overfitting

$$\mathbf{c}_{\text{rdc}} = (\hat{\mathbf{M}}^T\hat{\mathbf{M}} + \lambda\mathbf{R})^{-1}\hat{\mathbf{M}}^T\hat{\mathbf{a}} \quad (6)$$

- λR: parametric regularizer with strength λ

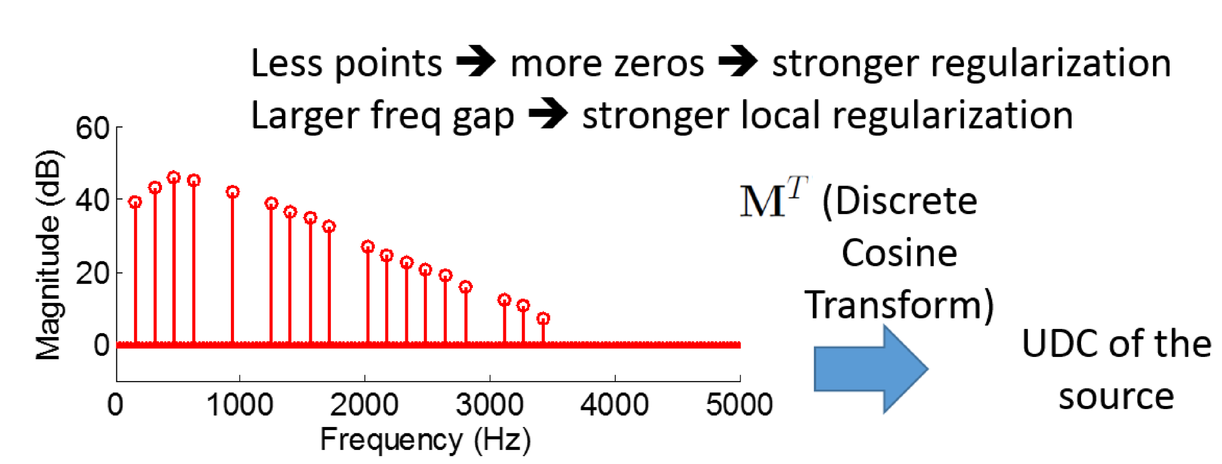
Proposed Cepstral Representations

▶ UDC: removing $(\hat{\mathbf{M}}^T\hat{\mathbf{M}})^{-1}$ in DC calculation

$$\mathbf{c}_{\text{udc}} = \hat{\mathbf{M}}^T\hat{\mathbf{a}} \quad (7)$$

$$= \mathbf{M}^T\tilde{\mathbf{a}} = N(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\tilde{\mathbf{a}} \quad (8)$$

- Eq. (8) shows that UDC is taking DCT on a sparse spectrum $\tilde{\mathbf{a}}$, which equals to $\mathbf{a}$ at selected spectral points and 0 otherwise.
- Eq. (8) also shows $\mathbf{c}_{\text{udc}}$ is the least square solution of approximating the sparse spectrum $\tilde{\mathbf{a}}$ with sinusoids.
- Zeros in $\tilde{\mathbf{a}}$ serve as a natural and locally adaptive regularizer.



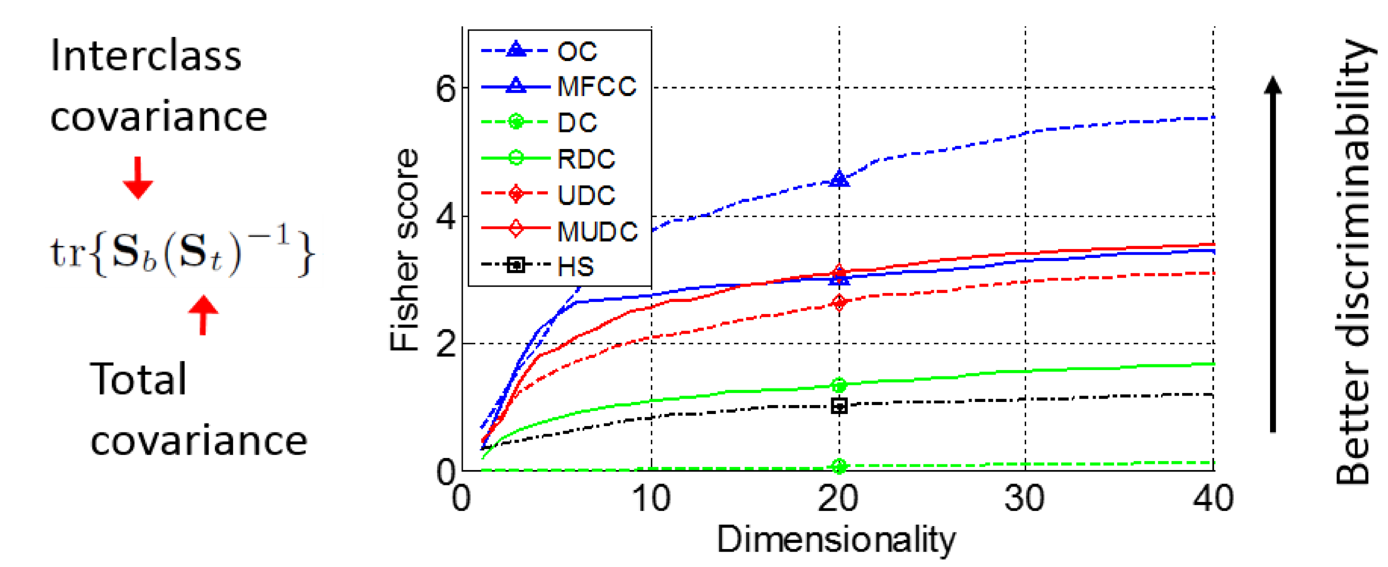
▶ Mel-scale UDC (MUDC)

- Let f be normalized mel-scale frequency 0.5mel(Hz)/mel(Fs/2).

Experiments

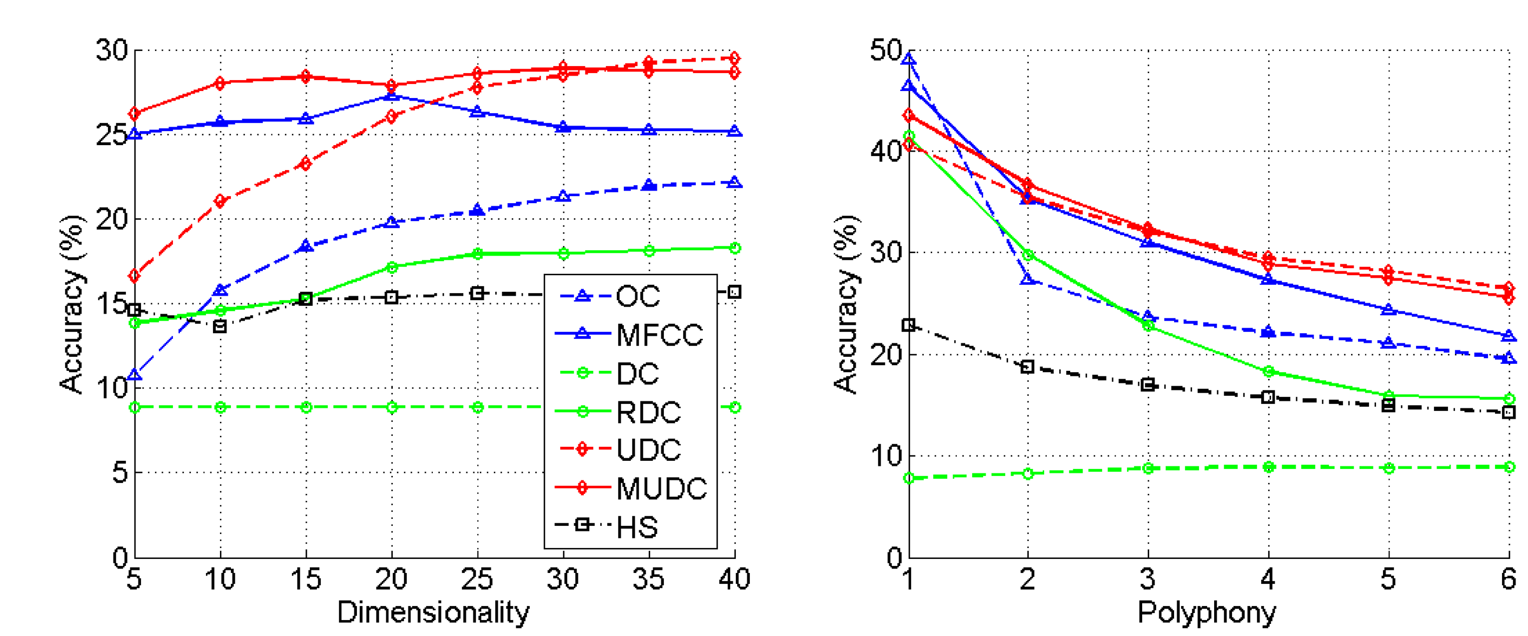
▶ Fisher score analysis of timbre representations

- 687 notes from 13 instruments of the University of Iowa dataset
- 5 frames randomly selected from the sustain part of each note
- Calculate OC and MFCC from full spectrum
- Detect pitch using YIN [1], then consider 50 harmonics as selected spectral points for DC, RDC, UDC, and MUDC



▶ Instrument recognition in polyphonic audio

- Train 13-class SVM with 687 notes from the U Iowa dataset
- Test on 5000 random chords mixed with notes from RWC database
- Detect ground-truth pitches using YIN [1] prior to mixing
- Calculate OC and MFCC from separated source spectrum, using a soft-masking-based separation method [2] with ground-truth pitches
- Calculate all the other features from first 50 harmonics of pitch



Conclusions

▶ UDC and MUDC use a more **natural and locally adaptive** regularizer to prevent overfitting the isolated spectral points.
▶ UDC and MUDC outperform the other representations in the task of instrument recognition in polyphonic audio mixtures.

References and Acknowledgement

[1] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music", *JASA*, 111: 1917-1930, 2002.
[2] Z. Duan and B. Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE J-STSP*, 5(6): 1205-1215, 2011.

This work was partially supported by NSF grant 1116384.