A NOVEL CEPSTRAL REPRESENTATION FOR TIMBRE MODELING OF SOUND SOURCES IN POLYPHONIC MIXTURES

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

[1]Department of Electrical and Computer Engineering, University of Rochester, USA.
[2]Department of Electrical Engineering and Computer Science, Northwestern University, USA.
[3]Institut Langevin, Université Paris Diderot - Paris 7, France.

ABSTRACT

We propose a novel cepstral representation called the uniform discrete cepstrum (UDC) to represent the timbre of sound sources in a sound mixture. Different from ordinary cepstrum and MFCC which have to be calculated from the full magnitude spectrum of a source after source separation, UDC can be calculated directly from isolated spectral points that are likely to belong to the source in the mixture spectrum (e.g., non-overlapping harmonics of a harmonic source). Existing cepstral representations that have this property are discrete cepstrum and regularized discrete cepstrum, however, compared to the proposed UDC, they are not as effective and are more complex to compute. The key advantage of UDC is that it uses a more natural and locally adaptive regularizer to prevent it from overfitting the isolated spectral points. We derive the mathematical relations between these cepstral representations, and compare their timbre modeling performances in the task of instrument recognition in polyphonic audio mixtures. We show that UDC and its mel-scale variant MUDC significantly outperform all the other representations.

Index Terms— Cepstrum, timbre, instrument recognition, polyphonic

1. INTRODUCTION

Timbre, also known as tone quality or tone color, plays an important role for humans in evaluating the aesthetics of a musical note articulation, in recognizing and discriminating sound events, and in tracking sound sources in polyphonic mixtures. Finding out good physical representations of timbre has been an active research topic for a long time. A good timbre representation would be useful in speaker identification and instrument recognition. It would also be useful for sound source tracking and separation.

Over the years, researchers have found that the rough spectral content and its temporal evolution characterizes timbre pretty well. Physical properties that quantify the spectral content include spectral centroid, skewness, kurtosis, spread, flatness, irregularity, and roll-off, among others [1]. Physical properties that quantify the temporal evolution of the spectral content include spectral flux, vibrato/tremolo rate and depth, and the attack/release time of the amplitude envelope [1]. Another category of representations assume the source-filter model of sound production, where the source (excitation) signal carries the pitch information and the frequency response of the resonance filter determines the timbre. The frequency response of the filter is invariant to pitch. Researchers have proposed different ways to represent the filter, some are in the time domain such as linear predictive coding (LPC) [2] and its perceptual modification PLP [3], while others are in the cepstrum domain [4] such as mel-frequency cepstral coefficients (MFCC) [5].

These above-mentioned timbre features have shown great success in sound synthesis, speech recognition, speaker and instrument identification, music genre classification, etc. However, they have a common limitation: they cannot model the timbre of a sound source in a mixture without resorting to source separation, because their calculation requires the whole signal/spectrum of the sound source. However, source separation is an extremely difficult problem.

In this paper we are interested in timbre features for sound sources that can be calculated from the mixture signal directly, without resorting to source separation. To simplify this problem, we assume the sources are harmonic sources and their pitches have been correctly estimated. It is noted that even in this case, source separation is a hard problem, due to overlapping harmonic issues and reconstruction of nonharmonic regions.

The harmonic structure feature (HS), proposed in [6], is defined as the relative log-amplitudes of the harmonics of the source. It can be calculated from the sound mixture directly without source separation, assuming the pitch is provided. It has been shown to successfully model the timbre of the sound source for source separation [6] and multi-pitch streaming [7]. However, it is only pitch-invariant within a narrow pitch range (say one octave) [6].

Discrete cepstrum (DC), proposed by Galas and Rodet [8], is a cepstral representation of a sound source that can be calculated from a sparse set of points of its spectrum. For harmonic sound sources, the frequencies are the (non-overlapping) harmonics. Therefore, like harmonic structure, it can be calculated for a sound source from the mixture signal directly without source separation. However, it has the issue that the reconstructed spectral representation overfits the sparse set of spectral points and oscillates a lot at other frequencies. Cappe et al. [9] identified this problem and imposed a regularization term to prevent the unwanted oscillations, and named the regularized representation the Regularized Discrete Cepstrum (RDC). Nevertheless, the strength of regularization is manually controlled, and is not easy to adapt for different frames of the signal. Both DC and RDC were proposed for spectral envelope reconstruction purposes and have never been tested in timbre discrimination experiments.

In this paper, we propose a new cepstral representation called uniform discrete cepstrum (UDC). Similar to DC and RDC, it is calculated from a sparse set of frequencies of the magnitude spectrum, hence can be calculated for each source from the mixture spectrum directly without source separation. The advantage of UDC is that it uses a natural and locally adaptive regularizer to prevent overfitting, hence is more robust in timbre modeling. In addition, its calculation is simpler than DC and RDC. In the experiments, we compare UDC and its mel-scale variant MUDC with other timbre representations, and show that they outperform others in a musical instrument recognition task from polyphonic audio.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. CALCULATION OF UDC AND MUDC

In this section, we describe how to calculate a UDC feature vector of a sound source from the mixture spectrum. Let $\mathbf{f} = [f_1, \cdots, f_N]^T$ and $\mathbf{a} = [a_1, \cdots, a_N]^T$ be the full set of normalized frequencies (Hz/Fs, Fs being the sampling frequency in Hz) and log-amplitudes (dB) of the mixture spectrum of discrete Fourier transform (DFT). Suppose $\hat{\mathbf{f}} = [\hat{f}_1, \cdots, \hat{f}_L]^T$ and $\hat{\mathbf{a}} = [\hat{a}_1, \cdots, \hat{a}_L]^T$ are the sparse subset of the spectral points that are likely to solely belong to the source we want to model[1], which we call the *observable spectral points* for the source. Then the UDC is calculated by

$$\mathbf{c}_{\text{udc}} = \hat{\mathbf{M}}^T \hat{\mathbf{a}}, \tag{1}$$

where

$$\hat{\mathbf{M}} = \begin{pmatrix} 1 & \sqrt{2}\cos(2\pi 1 \hat{f}_1) & \cdots & \sqrt{2}\cos(2\pi(p-1)\hat{f}_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \sqrt{2}\cos(2\pi 1 \hat{f}_L) & \cdots & \sqrt{2}\cos(2\pi(p-1)\hat{f}_L) \end{pmatrix}; \tag{2}$$

and p is the cepstrum order, i.e. the number of coefficients. The definition of Eq. (1) and (2) originates from the general concept of cepstrum, and will be discussed in Section 3.

If for $\hat{\mathbf{f}}$ in Eq. (2) we use normalized mel-scale frequencies instead of normalized frequencies, we obtain a mel-scale variant of UDC in Eq. (1), called MUDC, or $\mathbf{c}_{\text{mudc}}$. The normalized mel-scale frequencies is defined as 0.5mel(Hz)/mel(Fs/2), where $\text{mel(Hz)} = 2595\log_{10}(1 + \text{Hz} \times \text{Fs}/700)$;

The calculation of UDC and MUDC only requires the observable spectral points instead of the full separated spectrum of the source. For a harmonic source in an audio mixture, these observable spectral points can be the non-overlapping harmonic peaks given the pitch. It is noted that these points are not enough to reconstruct the spectrum of the source. Energy at overlapping harmonic peaks and non-peak regions need to be allocated to different sources in source separation as well.

3. RELATION TO OTHER CEPSTRAL REPRESENTATIONS

The concept of cepstrum [4] is to approximate (up to a scale) a log-amplitude spectrum $a(f)$ by a weighted sum of p sinusoids:

$$a(f) \approx c_0 + \sqrt{2}\sum_{i=1}^{p-1} c_i \cos(2\pi i f), \tag{3}$$

where the coefficients $\mathbf{c} = [c_0, c_1, \cdots, c_{p-1}]^T$ form a cepstrum of order p; f is the normalized frequency. By varying f, Eq. (3) forms a linear equation system, where the number of equations is the number of frequencies at which we make the approximation. A common approximation criterion is to minimize the Euclidean distance between both sides, which leads to the least squares solution of the coefficients.

It turns out that the *ordinary cepstrum (OC)* is the least square solution when we make the approximation at all the N frequency bins $\mathbf{f}$. There are in total N equations, which can be written in the matrix notation as:

$$\mathbf{a} = \mathbf{M}\mathbf{c}, \tag{4}$$

where

$$\mathbf{M} = \begin{pmatrix} 1 & \sqrt{2}\cos(2\pi 1 f_1) & \cdots & \sqrt{2}\cos(2\pi(p-1)f_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \sqrt{2}\cos(2\pi 1 f_N) & \cdots & \sqrt{2}\cos(2\pi(p-1)f_N) \end{pmatrix}, \tag{5}$$

consists of the first p columns of a discrete cosine transform (DCT) matrix. The least square solution of the coefficients is

$$\mathbf{c}_{\text{oc}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{a} = \frac{1}{N}\mathbf{M}^T\mathbf{a}, \tag{6}$$

where the last equality follows that the columns of $\mathbf{M}$ are orthogonal and all have a Euclidean norm of $\sqrt{N}$.

$\mathbf{c}_{\text{oc}}$ is calculated by approximating the full log-amplitude spectrum and it reconstructs a smoothed version of the spectrum. If the spectrum is warped into a mel-scale filterbank before the cepstrum calculation, then the cepstrum is the so called *mel-frequency cepstral coefficients (MFCC)*. Both OC and MFCC have been shown to perform well in timbre discrimination, when they are calculated from isolated recordings of sound sources [10]. However, from a mixture spectrum containing multiple sound sources as what we are interested in this paper, they cannot be calculated to represent the timbre of the sound sources without source separation.

There does exist a cepstral representation called *discrete cepstrum (DC)* proposed by Galas and Rodet [8] that can be calculated from only a sparse set of spectral points instead of the full spectrum. In fact, DC is defined as the least square solution of Eq. (3) when the approximation is made only at the L observable spectral points, i.e. the following system of L equations:

$$\hat{\mathbf{a}} = \hat{\mathbf{M}}\mathbf{c}, \tag{7}$$

where $\hat{\mathbf{M}}$ is given in Eq. (2). Its least square solution is

$$\mathbf{c}_{\text{dc}} = (\hat{\mathbf{M}}^T\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}^T\hat{\mathbf{a}}. \tag{8}$$

Since the approximation is only performed at the L observable spectral points, $\mathbf{c}_{\text{dc}}$ reconstructs a smooth curve that goes through the observable spectral points and ignores the other parts of the spectrum. When these points are harmonics of a source, this curve is a *spectral envelope* of the source spectrum. Representations of spectral envelopes are essential for sound synthesis and this was what DC was proposed for in [8]. However, it can also be used for timbre discrimination, although it has never been tested before.

Eq. (7) has L equations and p unknowns. One needs to make $p <= L$ to obtain unique solutions. However, this requirement is not always satisfied since the number of observable spectral points L of the target source may vary significantly in different time frames of the mixture spectrum. Furthermore, the matrix $\hat{\mathbf{M}}^T\hat{\mathbf{M}}$ is often poorly-conditioned due to the large frequency gap between some observable spectral points. This means that non-significant perturbations of the observable spectral points may cause large variations of the estimated coefficients. The reconstructed spectral envelope tends to overfit the observable spectral points of the source, while oscillating significantly at the other frequencies.

This problem of $\mathbf{c}_{\text{dc}}$ was identified by Cappé et al. in [9]. They then proposed a *regularized discrete cepstrum (RDC)* by introducing to the least square system a regularization term, which prefers solutions that reconstructs smoother spectral envelopes:

$$\mathbf{c}_{\text{rdc}} = (\hat{\mathbf{M}}^T\hat{\mathbf{M}} + \lambda\mathbf{R})^{-1}\hat{\mathbf{M}}^T\hat{\mathbf{a}}, \tag{9}$$

[1] In fact, $\hat{\mathbf{f}}$ need not to be a subset of frequency bins in Fourier analysis. They can be frequencies in between the bins, and $\hat{\mathbf{a}}$ can be the corresponding interpolated values. In this case, the first equality of Eq. (10) will be an approximation.

where $\mathbf{R}$ is a diagonal matrix derived from a particular kind of regularization; λ controls the tradeoff between the original least square objective and the regularization term.

The proposal of UDC and MUDC was inspired by DC. Their calculation also only uses the observable spectral points of the interested sound source, hence they can be calculated from the mixture spectrum directly. This is an advantage over OC and MFCC, which require source separation first. Furthermore, by comparing Eq. (2) with Eq. (5) we can see that $\hat{\mathbf{M}}$ is a sub-matrix (a subset of rows) of $\mathbf{M}$, corresponding to the L observable frequency bins. Therefore, we can rewrite Eq. (1) as

$$\mathbf{c}_{\text{udc}} = \mathbf{M}^T\tilde{\mathbf{a}} = N(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\tilde{\mathbf{a}}, \tag{10}$$

where $\tilde{\mathbf{a}}$ is a sparse log-amplitude spectrum of the same dimensionality with the full mixture spectrum $\mathbf{a}$. It takes values of $\mathbf{a}$ at the sparse observable spectral points, and zeros everywhere else. Eq. (10) tells us that $\mathbf{c}_{\text{udc}}$ is equivalent to calculating the scaled (by N) ordinary cepstrum of the sparse spectrum $\tilde{\mathbf{a}}$. It is the scaled least square solution of $\tilde{\mathbf{a}} = \mathbf{M}\mathbf{c}$. It is noted that $\tilde{\mathbf{a}}$ would not serve as a good separated spectrum of the source. It is too sparse and its reconstructed source signal would contain musical noise.

Comparing Eq. (1) and Eq. (8), we can see that $\mathbf{c}_{\text{dc}} = (\hat{\mathbf{M}}^T\hat{\mathbf{M}})^{-1}\mathbf{c}_{\text{udc}}$. Therefore $\mathbf{c}_{\text{udc}}$ is not the least square solution for $\hat{\mathbf{a}} = \hat{\mathbf{M}}\mathbf{c}$, as $\mathbf{c}_{\text{dc}}$ is. This means that the reconstructed smooth curve from $\mathbf{c}_{\text{udc}}$ will not go through the observable spectral points as close as that reconstructed from $\mathbf{c}_{\text{dc}}$. In fact, since $\mathbf{c}_{\text{udc}}$ is the least square solution of $\tilde{\mathbf{a}} = \mathbf{M}\mathbf{c}$, it also needs to fit the zero elements in the sparse spectrum $\tilde{\mathbf{a}}$. From another perspective, the zero elements in $\tilde{\mathbf{a}}$ actually serve as another kind of regularizer that prevents $\mathbf{c}_{\text{udc}}$ from overfitting the observable spectral points.

Compared with the parameterized, global regularizer in RDC, this regularizer in UDC is non-parametric, adaptive, and local. Its strength varies naturally with the number (which is $N - L$) and pattern of the observable spectral points. When L is small in some frames, the regularizer is stronger. When there a big gap between two adjacent observable spectral points, the zero elements in between form a straight line and prevent significant oscillations of the reconstructed smooth curve in this gap. Furthermore, the calculation of UDC and MUDC is simpler than RDC and DC. The latter involves matrix inversion and multiple matrix multiplications while the former is just one matrix multiplication. In the following sections, we perform experiments to show that UDC and MUDC indeed represent timbre of sound sources and outperform other cepstral representations in instrument recognition from polyphonic mixtures.

4. EXPERIMENT ON ISOLATED NOTE SAMPLES

In the first experiment, we compare the six above-mentioned cepstral representations (OC, MFCC, DC, RDC, UDC, and MUDC) and the harmonic structure feature (HS), all calculated from the spectra of isolated note samples. We want to show that the proposed UDC and MUDC indeed characterize the timbre of musical instruments.

The dataset we use is the University of Iowa musical instrument samples database [11], which contains isolated note samples of a collection of Western pitched instruments recorded in different pitches, dynamics, and performing styles. We selected in total 687 notes from 13 instruments: flute, oboe, Bb clarinet, bassoon, alto saxophone, trumpet, horn, tenor trombone, tuba, violin, viola, cello, and bass. These notes cover the full pitch range of each instrument, and are all played in mezzo forte (mf) dynamic. Notes of string instruments are played in the arco style (i.e., with a bow).

For each note, we randomly select five frames (length of 46ms) in the sustain part. We apply a hamming window on each frame and perform discrete Fourier transform with four-times zero padding to obtain its spectrum. The OC and MFCC features are then calculated from the whole log-amplitude spectrum of each frame. We use Dan Ellis's implementation [12] with a 40-band mel filter bank in calculating MFCC features. DC, RDC, UDC, MUDC, and HS features are calculated from the harmonic peaks of the spectrum. We use YIN [13] to detect the ground-truth pitch of the frame. Peaks that are within a quarter tone of a harmonic position is considered a harmonic peak. Only the first 50 harmonic positions are considered.

For each feature, we calculate the Fisher score [14] to quantify its discrimination power on instrument timbre:

$$\text{Fisher score} = \text{tr}\{\mathbf{S}_b(\mathbf{S}_t)^{-1}\}, \tag{11}$$

where $\mathbf{S}_b$ is the between-class scatter matrix which measures the scatterness of the representative points (the averages) of different classes, and $\mathbf{S}_t$ is the total scatter matrix which measures the scatterness of all the data points. Larger Fisher scores indicate better discrimination power hence better timbre modeling performance. Therefore, we prefer timbre features that give a large Fisher score.

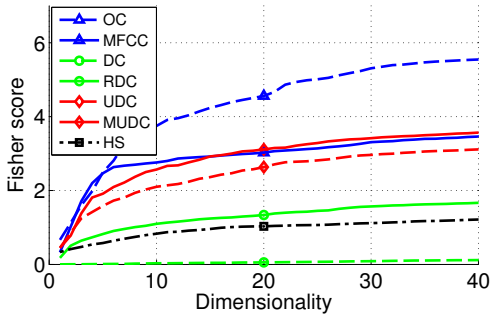

Fig. 1. Fisher score of the seven different features versus the dimensionality used in the features, calculated from 5 random frames of the sustain part of 687 isolated note samples of 13 Western instruments.

Figure 1 shows the Fisher scores calculated for different features versus dimensionality, i.e. the number of first coefficients used in the calculation. We can see that OC achieves the highest Fisher scores for all dimensionality and MFCC also achieves high scores. This is expected as they are calculated from the whole spectrum while the other features are calculated only from the harmonics. It is interesting to see that UDC and MUDC achieve Fisher scores comparable to MFCC. When the dimensionality is larger than 15, the Fisher score of MUDC even slightly exceeds MFCC. The gap between UDC and the other three features are very wide at all dimensionality. RDC and HS achieve similar Fisher scores while DC achieves the worst score. The bad performance of DC is expected due to its overfitting problem described in Section 3.

5. EXPERIMENT ON INSTRUMENT RECOGNITION FROM POLYPHONIC MIXTURES

We now compare the seven features on an instrument recognition task from polyphonic audio mixtures. We want to show advantages of the proposed UDC and MUDC over the other features on this task.

We still considered the 13 kinds of Western instruments in this experiment. We trained a multi-class SVM classifier using the LIBSVM package [15] on the features calculated from the 687 isolated

notes from the University of Iowa database described in Section 4. Again, five frames in the sustain part of each note were randomly selected, resulting in 3435 training vectors for each kind of feature. We normalized each dimension of the training feature vectors to the [-1, 1] range. We used a radial basis function (RBF) kernel and tuned the cost parameter C among $\{1, 10, 100, 1000, 10000\}$ for each feature. The best value was found using 5-fold cross validation on the training feature vectors when the dimensionality of 20 was used.

We tested the classifier using randomly mixed chords of polyphony from two to six, using isolated note samples from the RWC musical instrument dataset [16]. In total 1556 notes performed in mezzo forte without vibrato were selected from the 13 kinds of instruments. The notes of each kind of instrument were performed using three different brands of that instrument by three different players. The notes cover the full pitch range of the instrument. To generate a testing mixture of polyphony P, we first randomly chose without replacement P types of instruments. We then randomly chose a single note for each instrument, and a single frame in the sustain part of that note. We mixed the selected P frames with equal RMS values into a mixture frame. We used YIN [13] to detect the ground-truth pitch of each source before mixing. For each polyphony, we generated 1000 such mixtures.

For each source in each mixture, we calculated a timbre feature and classified it using the trained SVM. For OC and MFCC, the feature vector was calculated from the separated spectrum of the source using a soft-masking-based source separation method [17], which takes the ground-truth pitches as input. For DC, RDC, UDC, MUDC, and HS, the feature vector was calculated from the harmonic peaks of the source in the mixture spectrum, provided the ground-truth pitches. The percentage of correctly classified feature vectors over the total number of feature vectors is the classification accuracy. Since there are 13 instruments, the random classification accuracy would be roughly 8%, without considering the imbalance of the number of notes played by different instruments.

Figure 2 shows the average classification accuracies over 10 runs (1 run = data generation + training + testing) using different features versus the feature dimensionality. We can see that among all the seven features, MUDC achieves the highest accuracy at all dimensionality, and the accuracy does not change much with dimensionality. UDC's result is significantly better when the dimensionality is increased. MFCC also achieves high accuracy, however, it is sensitive to dimensionality. A two-sample t-test shows that MUDC achieves significantly higher average accuracy than MFCC at all dimensionality, at the significance level of 0.005.

Figure 3 further compares the seven different features on audio mixtures with different polyphony. For each feature and polyphony, the best dimensionality of the feature was used. Again, the figure shows the average results over 10 runs. From this figure, we can see that OC and MFCC achieve the best performance when polyphony is 1, which is in accordance with the results shown in Figure 1. The highest accuracy is about 50%, which sets the upper bound for all different polyphony settings in this cross-dataset instrument recognition experiment. For polyphony larger than 1, UDC and MUDC are again always the best features. For polyphony of 2, 3 and 4, MFCC performs almost as well as UDC and MUDC, despite that MFCC is more sensitive to feature dimensionality as shown in Figure 2. However, with the increase of polyphony, the gap between UDC/MUDC and MFCC becomes larger, indicating that the advantages of MUDC and UDC can be better shown for more complex audio mixtures, where satisfying source separation results for MFCC are more difficult to obtain. A two-sample t-test shows that MUDC outperforms MFCC significantly at all polyphony larger than 1 while UDC outperforms MFCC for all polyphony larger than 2, at the significance level of 0.005. Features of OC, RDC, and HS achieve better than chance but significantly lower accuracies, while DC, as expected, again achieves the chance accuracies. Classification here was performed in each single frame using a single type of feature. Combining results in different frames and using multiple features would improve the performance, but exceeds the scope of this paper.

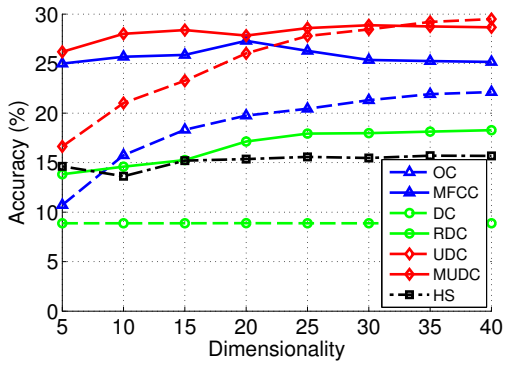

Fig. 2. Average instrument classification accuracy (over 10 runs) versus dimensionality of seven features, on 1000 random chords with polyphony of 4 in each run.

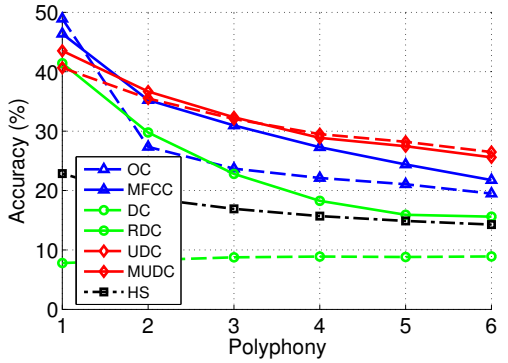

Fig. 3. Average instrument classification accuracy (over 10 runs) versus polyphony of audio mixtures. For each feature and polyphony, the best dimensionality was used.

6. CONCLUSIONS

We proposed a new cepstral representation called the uniform discrete cepstrum (UDC) and its mel-scale variant MUDC to characterize the timbre of sound sources in audio mixtures. Compared to ordinary cepstrum and MFCC, they can be calculated from the mixture spectrum directly without resorting to source separation. Compared to discrete cepstrum and regularized discrete cepstrum, they are easier to compute and have better discriminative power. We showed in experiments that they outperform the other five timbre features significantly in instrument recognition from polyphonic mixtures when the polyphony is high.

We thank reviewers for the valuable comments. Bryan Pardo was supported by the National Science Foundation grant 1116384.

7. REFERENCES

[1] Anssi Klapuri and Manuel Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] John Makhoul, "Spectral linear prediction: properties and applications," *IEEE Trans. Audio Speech Signal Processing*, vol. 23, pp. 283–296, 1975.

[3] Hynek Hermansky, "Perceptual linear predictive (plp) analysis of speech," *J. Acoust. Sos. Am.*, vol. 87, no. 4, 1990.

[4] Donald G. Childers, David P. Skinner, and Robert C. Kemerait, "The cepstrum, a guide to processing," in *Proc. IEEE*, October 1977, vol. 65, pp. 1428–1443.

[5] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, 1980.

[6] Zhiyao Duan, Yunggang Zhang, Changshui Zhang, and Zhenwei Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio Speech Language Processing*, vol. 16, no. 4, pp. 766–778, 2008.

[7] Zhiyao Duan, Jinyu Han, and Bryan Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE Trans. Audio Speech Language Processing*, vol. 22, no. 1, pp. 1–13, 2014.

[8] Thierry Galas and Xavier Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds," in *Proc. of International Computer Music Conference (ICMC)*, 1990, pp. 82–84.

[9] O. Cappé, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1995.

[10] Judy C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, vol. 105, pp. 1933–1941, 1999.

[11] Lawrence Fritts, "University of iowa musical instrument samples database," `http://theremin.music.uiowa.edu/MIS.html`.

[12] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.

[13] Alain de Cheveigné and Hideki. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[14] Quanquan Gu, Zhenhui Li, and Jiawei Han, "Generalized fisher score for feature selection," in *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

[15] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[16] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "Rwc music database: popular, classical, and jazz music databases," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[17] Zhiyao Duan and Bryan Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.