SPARSE MODELING FOR ARTIST IDENTIFICATION: EXPLOITING PHASE INFORMATION AND VOCAL SEPARATION

Li Su and Yi-Hsuan Yang
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
lisu@citi.sinica.edu.tw, yang@citi.sinica.edu.tw

ABSTRACT

As artist identification deals with the vocal part of music, techniques such as vocal sound separation and speech feature extraction has been found relevant. In this paper, we argue that the phase information, which is usually overlooked in the literature, is also informative in modeling the voice timbre of a singer, given the necessary processing techniques. Specifically, instead of directly using the raw phase spectrum as features, we show that significantly better performance can be obtained by learning sparse features from the negative derivative of phase with respect to frequency (i.e., group delay function) using unsupervised feature learning algorithms. Moreover, better performance is achieved by using singing voice separation as a pre-processing step, and then learning features from both the magnitude spectrum and the group delay function. The proposed system achieves 66% accuracy in identifying 20 artists from the artist20 dataset, which is better than a prior art by 7%.

1. INTRODUCTION

Singing voice is one of the most prominent characteristics in music. To model the vocal signal accurately, much of the effort has been focus on two topics: 1) extracting speech-related features of the audio signal [25, 27], and 2) separating human voice from the accompaniment [9,18,24, 29].

For feature extraction in speech processing, the phase-based features has been noticed with the application of speaker recognition or the reconstruction of intelligible voice [13, 21, 23, 26]. Instead of using phase only, these works adopted the *group-delay function* (the negative derivative of phase by frequency), sometimes also merged together with amplitude-based features. On the other hand, the application of phase on music has been limited mostly in signal-level, such as onset detection [2, 5, 16, 17] and pitch tracking [10]. For high-level musical concepts, like genre or the timbre of the vocal artist, phase information has been rarely discussed. Since singing timbre is closely related to characteristics of the speech signal, it is worth investigating the use of phase derivatives for artist identification from popular songs.

As for vocal separation, one remarkable development emerged recently is the sparse and low-rank matrix decomposition, also known as robust principal component analysis (RPCA) technique. It notices that the main melody and the accompaniment can be suitably regarded as the sparse and low-rank counterparts respectively in an audio spectrogram [15, 30], since the former involves only a few notes at a time, and the latter are typically contributed by repetition of metre and harmonic structure. This technique satisfies the requirement of artist identification, as such application needs an efficient algorithm to remove the accompaniment and preserve the vocal (mostly melody) information.

Recent years have witnessed progression of sparse modeling techniques [4, 6, 28], which allow for constructing a succinct representation of raw features as a combination of only a few *atoms* learned from an external data collection [22]. The resulting signal reconstruction has been shown robust to noise or corruptions of data. In consequence, sparse coding techniques have been applied in many fields, including MIR. For example, Yeh *et al.* [31] demonstrated accuracy on par with state-of-the-systems for genre classification using sparse features learned by sparse modeling techniques.

This paper will investigate the sparse modeling techniques for singing voice separation and unsupervised feature learning of group-delay functions. In what follows, we will provide the details of the proposed system in Section 2, followed by experimental evaluations in Section 3. We will discuss the main findings and limitations in Section 4 and conclude the paper in Section 5.

2. SYSTEM OVERVIEW

Figure 1 shows the flow diagram of the system we implemented for artist identification. The system makes use of a collection of songs, named the "training corpus," to build the audio dictionaries, which are used to compute the sparse representation of a querying audio signal. Prior to dictionary learning or sparse coding, frame-level features are extracted from the audio files. We consider both the *magnitude-based* and *phase-based* features derived from the short-time Fourier transform (STFT), by taking the log-magnitude spectrogram as the amplitude-based feature,

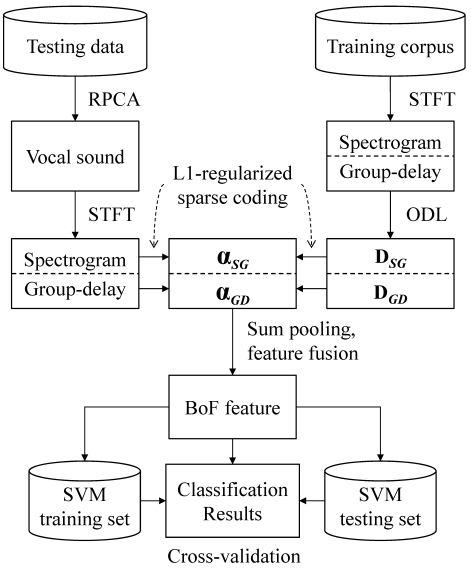

Figure 1. Proposed artist identification system.

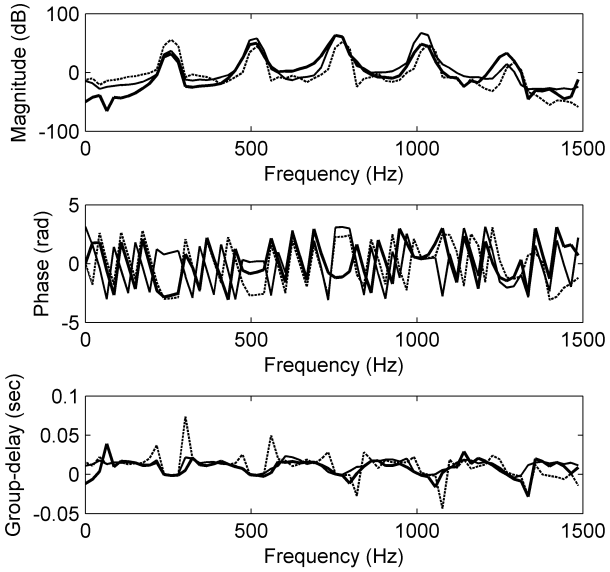

Figure 2. Examples of magnitude spectra, group-delay functions and phase profiles for tenor vocal /ah/ with fundamental frequencies at C4, selected from RWC Musical Instrument Sound Database [11]. Bold solid line: normal (less vibrato); thin solid line: vibrato; thin dashed line: falsetto.

and the group-delay as the phase-based feature.

Two types of dictionaries, $\mathbf{D}_{SG}$ and $\mathbf{D}_{GD}$, both of size $m \times k$, are respectively learned from the spectrograms and group-delay functions gathered from the training corpus using the online dictionary learning (ODL) algorithm (see Section 2.2). When there is a large overlap of artists between the training corpus and the querying songs, the generalizability of the learned dictionary might be limited (this is the so-called *transductive learning* setup). Therefore, it is preferable to use a training corpus that is representative enough, but is disjoint from the querying songs in the experiments.

As for testing, we first separate the query songs into singing voice and music accompaniment using RPCA technique, one of the state-of-the-art algorithms for singing voice separation [15]. Then, the log-magnitude spectrogram and group-delay functions are then encoded by $\mathbf{D}_{SG}$ and $\mathbf{D}_{GD}$ respectively by l_1-regularized sparse coding, engendering the codewords α_{SG} and α_{GD}. Each input feature is normalized to its Euclidean norm before sparse coding. After sparse coding, bag-of-frames (BOF) features are obtained by summing over all the frame-based features α_{SG} and α_{GD}, respectively, thereby creating a histogram of the cumulative term occurrence of the dictionary atoms [19]. Then we perform feature fusion by concatenating the two BOF features encoded from spectrograms and group-delay functions. Finally, the performance is evaluated by multi-class support vector classification in a cross-validation scheme.

2.1 Group Delay Function

Consider a general representation of short-time Fourier transform (STFT) of a time-domain signal $x(t)$:

$$\mathrm{S}_x^h(t,\omega) = \int_{-\infty}^{\infty} x(\tau)\, h^*(t-\tau)\, e^{-j\omega\tau} d\tau \quad (1)$$

$$= \mathrm{M}_x^h(t,\omega)\, e^{j\Phi_x^h(t,\omega)}, \quad (2)$$

where $\mathrm{S}_x^h(t,\omega) \in \mathbb{C}$ is the two-dimensional STFT representation on time-frequency plane, $h(t)$ is the window function, $\mathrm{M}_x^h(t,\omega)$ and $\Phi_x^h(t,\omega)$ of (1) are the amplitude and the phase of the STFT representation, respectively. By taking the natural logarithm of Eq. (2), we obtain the *log-magnitude spectrogram* in the real part and the *phase* the imaginary part:

$$\log \mathrm{M}_x^h(t,\omega) = \mathrm{Re}\left(\log \mathrm{S}_x^h(t,\omega)\right), \quad (3)$$

$$\Phi_x^h(t,\omega) = \mathrm{Im}\left(\log \mathrm{S}_x^h(t,\omega)\right). \quad (4)$$

Taking the negative derivative of phase (4) with respect to frequency, we have

$$-\frac{\partial \Phi_x^h(t,\omega)}{\partial \omega} = \mathrm{Re}\left(t - \frac{\mathrm{S}_x^{\mathcal{T}h}(t,\omega)}{\mathrm{S}_x^h(t,\omega)}\right), \quad (5)$$

where $\omega = 2\pi f$ is the angular frequency and $\mathcal{T}(\cdot)$ is the operator such that $\mathcal{T}h(t) = t \cdot h(t)$. The first term t denotes the current time, while the second term is defined as *group-delay function*. Detailed derivation procedures of group-delay function can be found in [1, 12]. In this work, the group-delay function is computed by the Time-Frequency Toolbox. [1]

To illustrate the effect of singing timbre on phase, in Figure 2 we show the spectral amplitudes, group-delay functions and phase profiles of three different examples (normal, vibrato and falsetto) of single vowel /ah/ sung by a tenor singer with the same fundamental frequencies C4. Since the fundamental frequencies of these three sounds are the same, the peaks in the amplitude spectra are mostly overlapped. At the spectral peak frequencies,

[1] http://tftb.nongnu.org/

the group-delay values of all three sounds are nearly zero. However, except for the spectral peaks, the group-delay function of falsetto sound is largely different from those of normal and vibrato sounds. The group-delay of falsetto is more peaky, possibly because of the occurrence of the transmission zeros (dips) evident from the magnitude spectrum, which mostly correspond to the spurious peaks seen in the group-delay function. We cannot observe this kind of spurious peaks from the magnitude spectra alone, due to the non-stationary nature of the signal under analysis (singing with vibrato), as well as the lack of sufficient frequency resolution. In contrast, group-delay functions better indicate the characteristics in frequency bands with small-energy.

2.2 Dictionary Learning and Sparse Coding

The atoms of a dictionary are typically learned from a large-scale training corpus. To overcome the difficulty of loading data into memory, the ODL algorithm is adopted [22]. ODL comes with a mini-batch mechanism that learns the dictionary incrementally by using a part of the training corpus in each update. Specifically, during the dictionary learning process, the atoms are updated for each input feature of the finite set of training signals $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$ through the following joint optimization problem:

$$\begin{aligned} &\hat{\mathbf{D}} = \underset{\mathbf{D}\in\mathcal{C}}{\operatorname{argmin}} \tfrac{1}{N}\sum_{i=1}^{N}\left(\tfrac{1}{2}\|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1\right), \\ &\forall j = 1, \ldots, k,\ \mathbf{d}_j^T\mathbf{d}_j \leq 1\,, \end{aligned} \tag{6}$$

where $\mathbf{y}_i \in \mathbb{R}^m$ is the i-th frame-level feature (column vector) of the input data (i.e., either $\log \mathrm{M}_x^h$ or Φ_x^h) from the training corpus, $\alpha_i \in \mathbb{R}^k$ is the codeword, $\mathbf{D} \in \mathbb{R}^{m\times k}$ is the dictionary, and the atom $\mathbf{d}_j$ is the jth column vector of $\mathbf{D}$. We can solve for $\mathbf{D}$ and α_i by minimizing one while keeping the other fixed [22]. The optimization of α involves a typical l_1-regularized sparse coding problem, which can be described as

$$\hat{\alpha_t} = \arg\min_{\alpha_t} \|\mathbf{x}_t - \mathbf{D}\alpha_t\|_2^2 + \lambda\|\alpha_t\|_1\,. \tag{7}$$

This problem has been well-studied in the machine learning and statistics fields, under different names such as the basis pursuit [4] or the lasso problem [28]. In this work, we use the open-source package SPAMS [2] and the LARS-lasso algorithm [6] for ODL and sparse coding, respectively. The parameter λ is set to $1/\sqrt{m}$ as in [6].

2.3 Source Separation

Given a monaural music signal, we first compute its $m \times n$ spectrogram $\mathbf{M}_x^h$ through STFT. Then, we separate the singing voice $\mathbf{E}$ (sparse components) from the music accompaniment $\mathbf{A}$ (low-rank components) by formulating the problem as the following RPCA problem,

$$\min_{\mathbf{A},\mathbf{E}:\ \mathbf{M}_x^h=\mathbf{A}+\mathbf{E}} \|\mathbf{A}\|_* + \mu\|\mathbf{E}\|_1\,, \tag{8}$$

[2] http://spams-devel.gforge.inria.fr/

Corpus	Original	Vocal	Accompaniment
USPOP	62.6%	65.9%	58.5%
USPOP2	61.7%	65.5%	56.6%
MIR-1K	55.9%	63.4%	53.4%
RWC	54.2%	59.9%	49.9%

Table 1. Average accuracy using log-magnitude spectrogram BOF features for various audio signals and training corpora. The vocal and accompaniment parts are separated by RPCA technique.

where $\|\cdot\|_*$ denotes the trace norm of a matrix (the sum of its singular values), $\|\cdot\|_1$ is the l_1 norm that denotes the sum of the absolute values of matrix entries, and μ is a positive weighting parameter that can be set to $1/\sqrt{\max(m,n)}$ as recommended in [3]. This algorithm is proven to be robust against gross errors and outliers, in comparison to its l_2-regularized counterpart, the well-known PCA algorithm. As Eq. (8) is convex, efficient algorithms such as accelerated proximal gradient (APG) and augmented Lagrange multipliers (ALM) [3, 20] can be employed to compute $\mathbf{A}$ and $\mathbf{E}$ in an iterative fashion. Open-source implementation of such solvers can be found from the Internet. [3]

3. EXPERIMENT

3.1 Dataset and Experimental Setup

We evaluate the performance of artist identification using the artist20 dataset [7], [4] which consists of six albums (1,413 songs in total) sung by 20 artists. For each song, a 30-second length audio signal with both vocal and music accompaniment is clipped for evaluation. The clips are sampled at 16 kHz. Most of the songs in the artist20 dataset can also be found in the uspop2002 dataset, [5] which contains over 7,000 Western Pop songs.

For computing the STFT, we use the Hanning window and try different values of the window size w (in terms of samples) and the hop factor h (the ratio of hop size and window size). For classifier training and testing, the l_2-regularized l_2-loss support vector classifier in LIBLINEAR [6] is employed for efficiency. To avoid album effect, a six-fold jack-knife cross-validation scheme is conducted. Each fold contains only one album from every artist. As there are many possible fold partitions, we perform the random partitions for ten times to get the average classification accuracy. Two-tailed t-test is also performed (over the ten six-fold partitions) to evaluate whether the performance difference between different methods or parameter settings is significant.

[3] http://perception.illinois.uiuc.edu/matrix-rank/

[4] http://labrosa.ee.columbia.edu/projects/artistid/

[5] http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

[6] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

SG		w/o GD	w/ GD, w = 512			w/ GD, w = 1024			w/ GD, w = 2048		
			h = 0.1	h = 0.2	h = 0.5	h = 0.1	h = 0.2	h = 0.5	h = 0.1	h = 0.2	h = 0.5
w = 1024	h = 0.1	62.9%	63.3%	63.6%	64.0%	**64.5%**	64.3%	63.8%	**64.6%**	64.3%	63.8%
	h = 0.2	63.1%	62.1%	62.9%	64.2%	**64.4%**	**65.0%**	63.8%	**65.4%**	**64.7%**	64.3%
	h = 0.5	63.4%	58.1%	59.6%	62.1%	61.9%	62.5%	62.4%	**64.9%**	64.5%	63.9%

Table 2. Comparison of average accuracy among SG BOF features and fused SG + GD BOF features under various window sizes and hop factors. Vocal audio signal and MIR-1K training corpus are used. Data in bold style are those which achieve significant improvement in comparison to non-fused features under the same parameter settings.

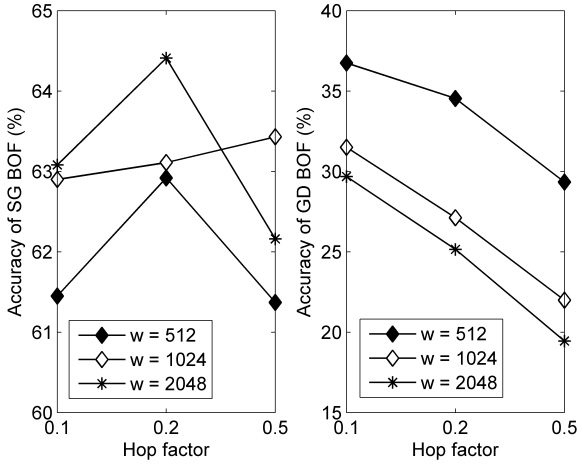

Figure 3. Average accuracy (in %) for SG BOF (left) and GD BOF (right) features constructed under various window sizes and hop factors. Vocal audio signal and MIR-1K training corpus are used.

3.2 Training Corpora

As discussed in Section 2, the size and the diversity of the training corpus are important to the generalizability of the learned dictionaries. We also have to pay attention to the possible overlap between the training corpus and the querying data (i.e., the dataset of the target classification problem; here artist20). To study the effect of training corpora, we compared the performance of artist identification using dictionaries learned from the following four training corpora: the whole uspop2002 dataset (USPOP), the uspop2002 dataset excluding overlap with artist20 (USPOP2), the MIR-1K dataset, and the whole vocal data in the RWC instrument dataset (RWC) [11]. MIR-1K contains 1,000 song clips extracted from 110 Chinese Pop songs released in karaoke format [14]; we use the vocal channel to train the dictionary. RWC vocal data contains five vowels (/a/, /i/, /u/, /e/, /o/) with various pitches and singing techniques sung by eight female and ten male singers, totaling more than 20,000 isolated notes. Among the four corpora, USPOP has the largest overlap with artist20. USPOP2 has no overlap but the music genre is the same (Western Pop). In contrast, the MIR-1K and RWC are considered more dissimilar from artist20.

As the first evaluation, we consider only BOF features computed from the log-magnitude spectrogram, using $w = 1024$ and $h = 0.5$. The size k of the dictionary is set to 1024. The evaluation result is shown in Table 1. The three columns of the table show the averaged accuracy using the original signal and the separated signals (vocal and accompaniment). By comparing the result of the four rows along the first column, we see that USPOP and USPOP2 lead to better accuracy comparing to MIR-1K and RWC. This is not surprising as the last two datasets are less similar with artist20. Although USPOP performs slightly better than USPOP2, the performance difference is not significant under the t-test.

3.3 Source Separation

By comparing the result of the three columns of Table 1, also it can be observed that using the separated vocal sound generally improves the classification accuracy, in comparison to using the original audio signals. In contrast, using the separated music accompaniment deteriorates the accuracy, which makes sense as the task emphasizes the vocal part of music. Two-tailed t-test shows that the improvement of Vocal over Original is significant ($p<0.001$, df=118). Moreover, when features are extracted from the vocal part, we see that the performance of using MIR-1K as the training corpus comes close to the case when USPOP is used (63.4% vs 65.9%).

To ensure the evaluation reported here is general enough, we use MIR-1K as the training corpus in the following experiments. Moreover, the features are computed from the separated vocal part of the song clips.

3.4 Incorporating Group Delay

Figure 3 shows the average accuracy of the BOF features computed from log-magnitude spectrogram (SG) and group-delay (GD) under various window sizes w and hop factors h. The sizes for the SG and GD dictionaries are both set to 1024. Note that we use different y-axes for the two features. As the figure shows, the performance of GD (20–35%) is generally much worse than the performance of SG (61–64%), possibly due to the highly noisy parts of the phase information. However, the result of GD is by no means random; the accuracies are significantly better than the random guess (whose accuracy is close to 5%). Moreover, we see a clear trend of the performance of GD with respect to w and h: better result is obtained by using smaller w and smaller h. In contrast, the performance of SG seems to be less sensitive to w and h.

We further experiment with the option of fusing the two types of features. The result is shown in Table 2, where we

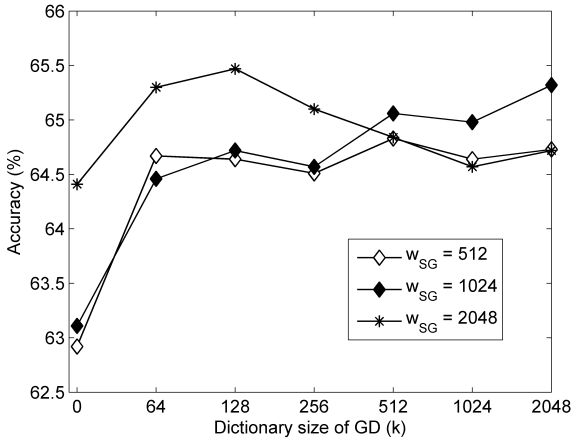

Figure 4. Average accuracies of different SG + GD BOF features by varying the dictionary size (k) and the window sizes of SG feature. Vocal audio signal and MIR-1K training corpus are used. As for GD feature, the window size is fixed to 1024. Hop factors for both SG and GD are 0.2.

compare the result without ('w/o') and with ('w/') using GD features with different values of w and h. We see the best result is obtained when SG (w=1024, h=0.2) and GD (w=2048, h=0.1) are fused, each using different values of w and h. Comparing to the case when GD is not used (i.e., the third column), the accuracy is improved from 63.1% to 65.4%, a significant improvement under the t-test ($p<0.05$, $df=118$). Actually, significant improvement is also observed for other cases, as indicated by the use of bold font weight in Table 2. This result shows that phase information is indeed useful for this task.

3.5 Influence of Dictionary Size

It has been shown that the performance of dictionary-based approach can be improved by using a larger dictionary [31]. As this is only verified on spectrogram-based features before, in this experiment we test the effects of the dictionary size k on the accuracy of artist identification using the group-delay features. Instead of using the GD features alone, we fuse it with SG features that is computed from a dictionary of size 1024, as this brings about better performance. Figure 4 shows the results as the dictionary size for GD ranges from 0 (no fusion) to 2048. When the window sizes w of SG are 512 or 1024, the performance gradually increases until reaching a plateau as we increase the dictionary size of GD. Generally speaking, using larger dictionary is also beneficial to phase-based features.

3.6 Comparison with the Existing Work

Finally, we compare our method with the GMM-based model proposed in [7], whose underlying frame-level feature representation is based on the early fusion of the classic MFCC and Chroma. Under a six-fold jack-knife cross-validation, the overall accuracy on artist20 was 56% for MFCC features and 59% for MFCC+Chroma features. Using the same fold partition and dictionaries learned from separated vocal parts of USPOP2, the proposed method reaches 66.0%, an improvement of 7%. Even if the less similar MIR-1K dataset is employed for dictionary learning, the classification accuracy reaches 65.5%.

4. DISCUSSION

In view of the source-filter model of voice, group-delay function contains the information on the phase-distortion behaviors of the vocal tract filter, which varies with the individual. Differently, the magnitude part contributes to the distribution of the resonance peaks on the time-frequency plane. Mathematically, the magnitude and the phase part of a STFT profile are strongly related, but they still have different characteristics [8]. Although the information provided by the magnitude part is prominent, the incorporation of group-delay usually better explains the characteristics of the target signal. There are many possibilities combining these two features, such as merging the two features at frame level [13], combining at BOF level, or in decision stage. Therefore, the way by which the features are combined and the relative weights of the two features are worthy for future study.

The use of singing voice separation generally improves the modeling of either the magnitude or phase information. However, the RPCA technique adopted in this work is not perfect. Under many cases the sparse counterpart of RPCA technique contains not only vocal sources but predominant instrument solo or the percussion sound. This issue might be partially solved by performing vocal detection first to exclude non-vocal segments.

5. CONCLUSION

In this paper, we have proposed a novel artist identification method based on sparse features learned from both the magnitude and phase parts of the spectrum. The features are computed from the separated vocal part of music signal using robust principal component analysis, in order to better model the characteristics of singing timbre. Our analysis shows that both singing voice separation and unsupervised feature learning are required steps for the features to be informative. Moreover, group-delay functions contain information complementing spectrogram. Evaluation on the artist20 dataset validates the effectiveness of the proposed phase feature. It is hoped that the present work can inspire more research towards the modeling of phase information of music signals, which might hold the promise of improving other MIR problems.

6. REFERENCES

[1] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the method of reassigment. *IEEE Trans. Signal Processing*, 43(5):1068–1089, 1995.

[2] E. Benetos and Y. Stylianou. Auditory spectrum-based pitched instrument onset detection. *IEEE Trans. Audio, Speech, Language Process.*, 18(8):1968–1977, 2010.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):1–37, 2011.

[4] S. S. Chen, D. L. Donoho, Michael, and A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scientific Computing*, 20:33–61, 1998.

[5] S. Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx06)*, 2006.

[6] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[7] D. Ellis. Classifying music audio with timbral and chroma features. In *ISMIR*, 2007.

[8] E. Chassande-Mottin F. Auger and P. Flandrin. On phase-magnitude relationships in the short-time fourier transform. *IEEE Signal Process. Lett*, 19(5):267–270, 2012.

[9] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Trans. Audio, Speech, Language Process.*, 18(3):638–648, 2010.

[10] M. Goto. A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings. In *IEEE ICASSP*, 2000.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *ISMIR*, 2003.

[12] S. Hainsworth, M. Macleod, S. W. Hainsworth, and M. D. Macleod. Time frequency reassignment: A review and analysis. Technical report, Cambridge University Engineering Department and Qinetiq, 2003.

[13] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Significance of the modified group delay feature in speech recognition. *IEEE Trans. Audio, Speech, Language Process.*, 15(1):190–202, 2007.

[14] C.-L. Hsu and J.-S. R. Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Trans. Audio, Speech, Language Process.*, 18(2):310–319, 2010.

[15] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *Proc. IEEE ICASSP*, 2012.

[16] S. Abdallah C. Duxbury M. Davies J. P. Bello, L. Daudet and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Trans. Speech, Audio Process.*, 13(5):1035–1047, 2005.

[17] A. Lacoste and D. Eck. A supervised classification algorithm for note onset detection. In *EURASIP Journal on Advances in Signal Processing*, 2007.

[18] M. Lagrange, A. Ozerov, and E. Vincent. Robust singer identification using melody enhancement and uncertainty learning. In *ISMIR*, 2012.

[19] M. Levy and M. Sandler. Music information retrieval using social tags and audio. *IEEE Trans. Multimedia*, 11(3):383–395, 2009.

[20] Z. Lin and et al. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Technical Report UILU-ENG-09-2214, 2009.

[21] L. Liu, J. L. He, and G. Palm. Effects of phase on the perception of intervocalic stop consonants. *Speech Communication*, 22:403–417, 1997.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, pages 689–696, 2009.

[23] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan. The delta-phase spectrum with application to voice activity detection and speaker recognition. *IEEE Trans. Audio, Speech, Language Process.*, 19(7):2026–2038, 2011.

[24] A. Mesaros, T. Virtanen, and A. Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *ISMIR*, 2007.

[25] T. L. Nwe and H. Li. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Trans. Audio, Speech, Language Process.*, 15(2):519–530, 2007.

[26] K. K. Paliwal and L. D. Alsteris. Further intelligibility results from human listening tests using the short-time phase spectrum. *Speech Communication*, 48:727–736, 2006.

[27] L. Regnier and G. Peeters. Combining classification based on local and global features: application to singer identification. In *DAFx-11*, pages 127–134, Sep. 2011.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statistical Soc.*, 58:267–288, 1996.

[29] W.-H. Tsai and H.-P. Lin. Background music removal based on cepstrum transformation for popular singer identification. *IEEE Trans. Audio, Speech, Language Process.*, 19(5):1196–1205, 2011.

[30] Y.-H. Yang. On sparse and low-rank matrix decomposition for singing voice separation. In *ACM MM*, pages 757–760, 2012.

[31] C.-C. M. Yeh and Y.-H. Yang. Supervised dictionary learning for music genre classification. In *ACM ICMR*, 2012.