ICA-FX Features for Classification of Singing Voice and Instrumental Sound

Tat-Wan Leung and Chong-Wah Ngo
Department of Computer Science
City University of Hong Kong
{csltw,cwngo}@cs.cityu.edu.hk

Rynson W. H. Lau
Department of CEIT
City University of Hong Kong
rynson@cs.cityu.edu.hk

Abstract

This paper describes a new approach in locating the segments of singing voice in pop musical songs. Initially, GLR distance measure is employed to temporally detect the boundaries of singing voices and instrumental sounds. ICA-FX is then adopted to extract the independent components of acoustic features for SVM classification. Experimental results indicate that ICA-FX can improve the classification performance by significantly reducing the independent components that are not related to class label information.

1. Introduction

Content-based musical information retrieval (MIR) has recently attracted numerous research attention due to its potential commercial applications in Internet musical search [2] and personalized musical retrieval system [6]. Broadly, we can classify the research efforts in MIR to six different categories as shown in Table 1. In general, the modeling of polyphonic signals is more difficult than monophonic signals since the former involves multi-dimensional note representation and source separation. Similarly, the content analysis of symbolic representation (*e.g.,* MIDI) is usually more straightforward than waveform representation (*e.g.,* MP3). To date, most people believe that the problems in Type-I category are solvable although not very interesting in practice, while the issues from types IV to VI are hard-to-solve but have great commercial potential [10].

Typical challenge in MIR, as indicated in Table 1, is in the case when monophonic signal (*e.g.,* humming) is used to match and retrieve polyphonic signals. The fundamental problems involved, for instance in Type-VI category, are two twofold: 1) how to effectively locate segments that contain singing voices in a song; 2) how to separate singing voices from background music. Once these problems are tackled, the matching between humming and the extracted singing voices can be solved by the algorithms like dynamic time warping or recurrent neural network [5].

This paper investigates the issues of locating singing voices in pop musical songs, particularly for Type-VI category. The major components in our approach include:

Type	Query	Database	Status
I	M	M	Mostly solved
II	P	P (MIDI)	Partially solved
III	P	P (Raw)	Partially solved
IV	M	P (MIDI)	Open problem
V	M	P (Raw)	Open problem
VI	M	P (Pop song, Raw)	Open problem

M: monophonic, P: polyphonic

Table 1. Research in MIR

- GLR distance measure is used to temporally detect the borders of singing voices and instrumental sounds for audio segmentation;
- A supervised version of ICA (independent component analysis) called ICA-FX [8] is used to extract independent components of audio segments while reducing the feature dimensions of acoustic feature vectors;
- Support Vector machine (SVM) is employed for pattern training and classification of singing voices and instrumental sounds.

The segmented singing voices from pop musical songs obviously is not "pure" singing voices, but rather a mixture of singing voices (SV) and instrumental sounds (IS). As a consequence, direct extraction of acoustic features from singing voices for classification may not give satisfactory results since the singing is actually mixed with IS. In this paper, we adopt ICA-FX to tackle this problem by extracting independent components from audio segments, while reducing feature dimensions. ICA-FX, unlike PCA (principle component analysis), attempts to maximize the mutual information between features and class information by assuming that the input features are mixed of N independent non-Gaussian sources. This assumption fits the targeted problem nicely since the extracted features from pop songs are actually a mixture of different sound sources. The extracted independent components by ICA-FX, theoretically, should outper-

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

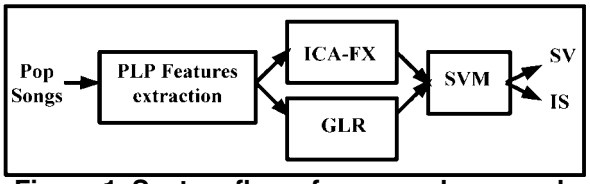

Figure 1. System flow of proposed approach

form the classification rate of the original feature space, if some of the independent components that are not related to class information can be successfully eliminated prior to classification.

2. Related Works

The issue of locating singing voices in musical songs has also been addressed in [1, 7]. In [1], a hidden Markov model (HMM) is employed for simultaneous segmentation and classification of audio signals. The acoustic features extracted include PLP coefficients and PPF vector generated by a neural network. PPF vector did not give good classification results as indicated in [1]. In this paper, we show that ICA-FX features extracted from PLP coefficients can outperform original PLP even though only tiny feature attributes are used.

In [7], a vocal detector is designed to analyze the harmonic energy in a specific frequency range. This is based on the assumption that the majority of energy in singing voices fall between 200Hz and 2000Hz. Nevertheless, this approach can achieve only approximately 55% of accuracy due to the fact that the energy of certain instrumental sounds like drums are also fallen in the range of interest.

Other related works include [5, 9] where the signal separation is carried out directly in frequency domain. In [5], ICA is used to analyze both stereo and monaural songs. Because the classification of SV and IS is not addressed, the results of ICA on the segments that contain only pure instrumental sounds are not analytical. In [9], a statistical model is developed for signal separation. As shown in their experiments, because the extracted pitch contours of singing voices are distorted by pure instrumental sounds, the precise matching between query and pop songs is not easy.

3. Overview of Proposed Approach

Figure 1 shows the flow of our proposed approach. Initially, pop songs are input to the system and PLP acoustic features are extracted. On one hand, the PLP features are used to locate the boundaries of SV and IS by GLR distance measure. On the other hand, ICA-FX is employed to transform and reduce the feature dimensions of PLP. The segmented audio signals, as well as the extracted ICA-FX features, are then used for pattern classification by SVM.

4. PLP Feature Extraction

Perceptual linear predictive coding (PLP) is the extent of linear predictive coding (LPC). Unlike LPC, PLP takes into account not only vocal tract, but also three concepts from psychophysics of hearing The audio features we use are 12^{th} order PLP cepstrum coefficients plus deltas and double deltas from each time frame. The delta and double delta are utilized to measure the velocity and acceleration of cepstrum changes respectively. The extracted PLP feature of a frame is represented as a 39-dimensional vector.

5. Audio Segmentation by GLR

Our aim is to detect the boundaries of singing voices and background instrumental sounds. As long as most boundary points are correctly detected, the results of classification will not be seriously affected even if it causes over-segmentation. The segmentation is achieved by comparing the GLR (generalized likelihood ratio) distance [4] of two neigbouring audio segments and then detecting the possible change points. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ as a sequence of PLP features along the temporal dimension, where $\mathbf{x}_i$ corresponds to the PLP of i^{th} frame. Two hypotheses are considered for boundary detection at i^{th} frame:

- H_0: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \sim \mathcal{N}(\mu, \Sigma)$
- H_1: $\mathbf{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_i\} \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbf{X}_2 = \{\mathbf{x}_{i+1}, \ldots, \mathbf{x}_n\} \sim \mathcal{N}(\mu_2, \Sigma_2)$

where $\mathbf{X}$ with mean vector μ and covariance matrix Σ is assumed to be generated by a mixture of Gaussian. The likelihood ratio between H_0 and H_1 is defined by

$$\lambda = \frac{L(\mathbf{X}, \mathcal{N}(\mu, \Sigma))}{L(\mathbf{X}_1, \mathcal{N}(\mu_1, \Sigma_1)) L(\mathbf{X}_2, \mathcal{N}(\mu_2, \Sigma_2))} \quad (1)$$

where $L(\mathbf{X}, \mathcal{N}(\mu, \Sigma))$ is the likelihood of $\mathbf{X}$ given model $\mathcal{N}(\mu, \Sigma)$. By Eqn (1), the GLR distance is computed as

$$\text{GLR Distance} = -\log \lambda \quad (2)$$

If there is a change of audio signals at frame i, the product of $L(\mathbf{X}_1, \mathcal{N}(\mu_1, \Sigma_1))$ and $L(\mathbf{X}_2, \mathcal{N}(\mu_2, \Sigma_2))$ will be greater than $L(\mathbf{X}, \mathcal{N}(\mu, \Sigma))$. Thus, the value of GLR distance at frame i should be high.

Figure 2 illustrates the computation of GLR distance by given the audio frames. $\mathbf{X}_1$ and $\mathbf{X}_2$ correspond to the PLP features in Win_1 and Win_2. The size of Win_1 and Win_2 is equal and is set to 0.5 second (approximately 32 frames). Both windows are sided by one frame after each GLR distance computation. A median filter is employed to remove high frequency noises of GLR distance values for robustness purpose. The boundary points for audio segmentation are then detected by locating the local maximum points of distance values, as shown in Figure 2.

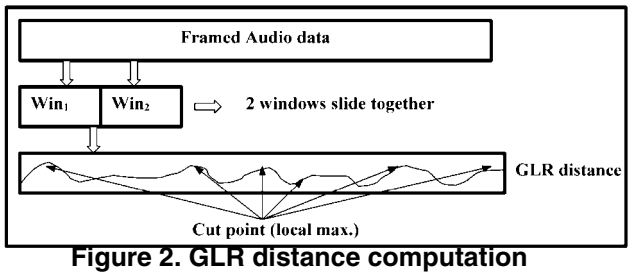

Figure 2. GLR distance computation

6. Feature Transformation by ICA-FX

We adopt ICA-FX [8] for the dimensionality reduction of PLP features. ICA-FX, in contrast to standard ICA, belongs to supervised learning and aims to maximize the mutual information between features and class label. Two sets of independent components will be generated. One set which, ideally, carries no information about class label will be discarded for classification.

Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ as PLP feature vector and $\mathbf{c}$ as class label. According to Fano's inequality [3], the lower-bound probability error of estimating $\mathbf{c}$ given $\mathbf{x}$ is

$$P_E(\mathbf{c}|\mathbf{x}) \geq \frac{H(\mathbf{c}|\mathbf{x}) - 1}{\log N_c} = \frac{H(\mathbf{c}) - I(\mathbf{x};\mathbf{c}) - 1}{\log N_c} \qquad (3)$$

where N_c is the number of classes, H and I are, respectively, the entropy and mutual information. In Eqn (3), the problem of minimizing $P_E(\mathbf{c}|\mathbf{x})$ is equivalent to maximizing $I(\mathbf{x};\mathbf{c})$. Nevertheless, there is no transformation that can increase $I(\mathbf{x};\mathbf{c})$. The aim of ICA-FX is to extract the feature $\hat{\mathbf{f}} = [f_1, f_2, \ldots, f_m]^T$ from a new transformed feature $\mathbf{f} = [f_1, \ldots, f_m, \ldots, f_n]^T$, where $m < n$, such that $I(\hat{\mathbf{f}};\mathbf{c})$ is as close as $I(\mathbf{x};\mathbf{c})$. Specifically, we have

$$I(\hat{\mathbf{f}};\mathbf{c}) \leq I(\mathbf{f};\mathbf{c}) = I(\mathbf{W}\mathbf{x};\mathbf{c}) \qquad (4)$$

where $\mathbf{W}$ is a $n \times n$ nonsingular matrix that transforms $\mathbf{x}$ to a new feature space $\mathbf{f}$ that contains $\hat{\mathbf{f}}$ as one of its subspaces.

The feature vector $\mathbf{x}$ can be considered as a linear combination of N independent non-Gaussian sources $\mathbf{s} = [s_1, s_2, \ldots, s_n]^T$. Denote $\mathbf{A}$ as a $n \times n$ mixing matrix and $\mathbf{b}$ as a $n \times 1$ vector, we have

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{b}\mathbf{c} \qquad (5)$$

The transformed feature $\mathbf{f}$ is obtained by unmixing $\mathbf{x}$. Let $\mathbf{u} = [u_1, u_2, \ldots, u_n]$ as the independent components and $\mathbf{v}$ as a $n \times 1$ vector. The feature $\mathbf{f}$ is computed by

$$\mathbf{f} = \mathbf{W}\mathbf{x} = \mathbf{u} - \mathbf{v}\mathbf{c} \qquad (6)$$

To reduce $\mathbf{f}$ by $n - m$ dimensions and extract $\hat{\mathbf{f}}$, the elements of $\mathbf{v}$ from $m + 1$ to n positions are set to zero. In this way, the feature vector $[f_{m+1}, \ldots, f_n]$ is independent of class label information and thus can be discarded.

To estimate $\mathbf{W}$, an iterative learning process based on maximum likelihood estimation is derived

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \alpha_1\{\mathbf{I} - \varphi(\mathbf{u})\mathbf{f}^T\}\mathbf{W}^{(t)} \qquad (7)$$

$$\mathbf{v_a}^{(t+1)} = \mathbf{v_a}^{(t)} - \alpha_2\varphi(\mathbf{u_a})\mathbf{c} \qquad (8)$$

where $\varphi(u_i) = -\frac{dp_i(u_i)}{du_i} \times \frac{1}{p_i(u_i)}$, $\mathbf{v_a}$ and $\mathbf{u_a}$ are respectively the sub-vectors of $\mathbf{v}$ and $\mathbf{u}$ without the elements from $m + 1$ to n. The probability of u_i, $p_i(u_i)$, is modeled by Laplace distribution. $\mathbf{I}$ is an $n \times n$ identity matrix and α_1 and α_2 are learning rates. The learning process terminates when $\Delta\mathbf{W} = \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}$ closes to zero matrix. Once the optimal solution for $\mathbf{W}$ is found, the ICA-FX feature $\hat{\mathbf{f}}$ which corresponds to the first m elements of $\mathbf{W}\mathbf{x}$ is extracted.

7. Support Vector Machine

Given the detected boundary points based on GLR distance measure, we want to classify each audio segment according to their acoustic features. The is done by classifying each frame in an audio segment based on the extracted ICA-FX features. The classification of a segment is based on the principle of winner-take-all. In other words, if the majority of frames in a segment vote for a particular class, the segment is regarded as belonging to that class.

We employ *C*-Support Vector Classification (*C*-SVC) for the pattern training and classification of audio frames. We use RBF (radial basis function) as the kernel function to map input vectors into higher dimensional feature space. SVM usually suffers from the slow training and testing speed especially for large data set such as audio frames. By using ICA-FX features (*e.g.,* 1-dimension), instead of original PLP features (39-dimension), both training and testing time can be drastically sped up, without significantly degrading the classification performance.

8. Experiments

To test the proposed approach, 5 pop musical songs are used for training, while 25 pop musical songs for testing. The audio data are down-sampled to 8kHz sample frequency. A $1 - 0.95z^{-1}$ filter is applied to pre-emphasize the audio data, and to complement the energy decrease in the high frequency bands of singing voices. We set the frame size as 16 ms with a 32 ms hamming window. All the songs are manually labeled "by ears" to indicate the boundaries of singing voices and instrumental sounds.

8.1. Classification Accuracy

Table 2 shows the classification rate of using PLP features and GLR distance measure. The average accuracy is 75%. Because GLR can cause over-segmentation of audio signals, we use the following measure to ensure fairness:

$$\text{Accuracy}_i = \frac{\sum_{seg_j \in \mathbf{c}_i} |seg_j|}{|\mathbf{c}_i|} \qquad (9)$$

where Accuracy$_i$ is the classification accuracy of class i, while $|\mathbf{c}_i|$ is the number of ground-truth frames in class i, and $|seg_j|$ is the number of frames in j^{th} segment.

	IS	SV
Instrumental Sound (IS)	72%	22%
Singing Voice (SV)	28%	78%

(Confusion Matrix)

Table 2. Classification by PLP features

The classification results based on ICA-FX features are shown in Table 3. At the frame-level, the classification accuracy of audio frames (without GLR) is tested. While at the segment-level, the evaluation based on Eqn (9) is experimented. As indicated in Table 3, when the feature dimension is reduced from 39 to 1 by ICA-FX, we can still achieve reasonably good performance in locating singing voices. The overall classification reaches the highest when the feature dimension equals to 6. Experimental results show that further increment in feature dimension only degrade the classification performance. This is probably due to the fact that the independent components after the sixth dimension are irrelevant to class information, and thus deteriorate the performance. Notice that when the dimension of ICA-FX features reaches 6, the classification rate is better than the original 39-dimensional PLP features.

Feature Dimension	Frame level		Segment level	
	IS	SV	IS	SV
1	62.56%	70.60%	64.68%	72.45%
2	62.73%	70.85%	64.97%	72.45%
3	63.52%	71.80%	66.67%	74.53%
4	64.81%	72.15%	67.12%	75.05%
5	67.10%	73.58%	70.90%	77.65%
6	68.28%	75.78%	73.14%	80.04%

Table 3. Classification results by ICA-FX

As shown in Table 3, the improvement of segment-level classification over frame-level classification is approximately 1% to 4%. This indicates that the winner-take-all strategy that we adopt for the classification of GLR partitioned audio segments is useful in getting rid of outliers and noise frames within a segment.

8.2. Performance Comparison

Table 4 compares the classification rate of features extracted by ICA-FX and PCA. As shown in the table, ICA-FX constantly outperforms PCA in all cases. The main reason is PCA seeks the directions in feature space that best represent the training data in a sum-square-sense. ICA-FX, in contrast, seeks the directions that are most independent from each other, while maximizing the mutual information between training data and class label information.

Feature Dimension	ICA-FX		PCA	
	IS	SV	IS	SV
1	64.68%	72.45%	57.50%	68.74%
2	64.97%	72.45%	58.21%	69.34%
3	66.67%	74.53%	59.24%	70.56%
4	67.12%	75.05%	61.62%	72.39%
5	70.90%	77.65%	63.99%	73.04%
6	73.14%	80.04%	64.33%	75.51%

Table 4. Comparison of ICA-FX and PCA

9. Conclusions

We have presented a new approach in classifying SV and background IS of popular musical songs. Through experiments, we show that ICA-FX is an effective tool in extracting discriminant audio features, while GLR distance measure, in addition to audio segmentation, is useful in isolating outliers within audio segments. Based the classified SV segments, our next step is to develop approach to extract "pure singing voices" up to certain orders of harmonicity directly in frequency domain for content-based pop song retrieval.

Acknowledgements

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7001546).

References

[1] A. L. Berenzweig and D. P. W. Ellis. Locating singing voice segments within musical signals. In *IEEE Workshop on application of Signal Processing to Audio and Acoustic*, 2001.

[2] W. Chai and B. Vercoe. Melody retrieval on the web. In *Proceedings of ACM/SPIE Multimedia Computing and Networking Conference*, Jan 2002.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[4] P. Delacourt, D. Kryze, and C. J. Wellekens. Speaker-based segmentation for audio data indexing. In *ESCA Workshop on Accessing Information in Audio Data*, 1999.

[5] Y. Feng, Y. Zhuang, and Y. Pan. Popular song retrieval based on singing matching. In *Int. Conf. on Music Information Retrieval*, 2002.

[6] K. Hoashi, K. Matsumoto, and N. Inoue. Personalization of user profiles for content-based music retrieval based on relevance feedback. In *ACM Multimedia Conf.*, 2003.

[7] Y. E. Kim and B. Whitman. Singer identification in popular music recordings using voice coding features. In *Int. Conf. on Music Information Retrieval*, 2002.

[8] N. Kwak and C.-H. Choi. Feature extraction based on ica for binary classification problems. *IEEE Trans. on Knowledge and Data Engineering*, 15(6):1374–1388, January 2003.

[9] T. W. Leung, C. W. Ngo, and W. H. Lau. Singing voice extraction from polyphonic pop song. In *Int. Conf. on Intelligent Multimedia Computing and Networking*, 2003.

[10] C. Yang. Efficient acoustic index for music retrieval with various degrees of similarity. In *ACM Multimedia Conf.*, 2002.