# SINGING VOICE SEPARATION USING SPECTRO-TEMPORAL MODULATION FEATURES

**Frederick Yen** **Yin-Jyun Luo**
Master Program of SMIT
National Chiao-Tung University, Taiwan
{fredyen.smt01g,fredom.smt02g}
@nctu.edu.tw

**Tai-Shih Chi**
Dept. of Elec. & Comp. Engineering
National Chiao-Tung University, Taiwan
tschi@mail.nctu.edu.tw

## ABSTRACT

An auditory-perception inspired singing voice separation algorithm for monaural music recordings is proposed in this paper. Under the framework of computational auditory scene analysis (CASA), the music recordings are first transformed into auditory spectrograms. After extracting the spectral-temporal modulation contents of the time-frequency (T-F) units through a two-stage auditory model, we define modulation features pertaining to three categories in music audio signals: *vocal*, *harmonic*, and *percussive*. The T-F units are then clustered into three categories and the singing voice is synthesized from T-F units in the vocal category via time-frequency masking. The algorithm was tested using the MIR-1K dataset and demonstrated comparable results to other unsupervised masking approaches. Meanwhile, the set of novel features gives a possible explanation on how the auditory cortex analyzes and identifies singing voice in music audio mixtures.

## 1. INTRODUCTION

Over the past decade, the task of singing voice separation has gained much attention due to improvements in digital audio technologies. In the research field of music information retrieval (MIR), separated vocal signals or accompanying music signals can be of great use in many applications, such as singer identification, pitch extraction, and music genre classification. During the past few years, many algorithms have been proposed for this challenging task. These algorithms can be categorized into unsupervised and supervised approaches.

The unsupervised approaches do not contain any training mechanism in the algorithms. For instance, Durrieu et al. used a source/filter signal model with non-negative matrix factorization (NMF) to perform source separation [5] and Fitzgerald et al. used median filtering and factorization techniques to separate harmonic and percussive components in audio signals [7]. Some other unsupervised methods considered structural characteristics of vocals and music accompaniments in several domains for separation. For example, Pardo and Rafii proposed REPET which views the accompaniments as repeating background signals and vocals as the varying information lying on top of them [16]. Tachibana et al. proposed the separation technique, HPSS, to remove the harmonic and percussive instruments sequentially in a two-stage framework by considering the nature of fluctuations of audio signals [19]. Huang et al. used RPCA to present accompaniments in low-rank subspace and vocal in sparse representation [8]. In addition, some unsupervised CASA-based systems were proposed for singing voice separation by finding singing dominant regions on the spectrograms using pitch and harmonic information. For instance, Li and Wang proposed a CASA system obtaining binary masks using pitch-based inference [13]. Hsu and Jang extended the work and proposed a system for separating both voiced and unvoiced singing segments from the music mixtures [9]. Although training mechanisms were seen in these two systems, they were only for detecting voiced and unvoiced segments, but not for separation.

In contrast, there were approaches based on supervised learning techniques. For example, Vembu et al. used vocal/non-vocal SVM and neural-network (NN) classifiers for vocal-nonvocal segmentation [20]. Ozerov et al. used a vocal/non-vocal classifier based on Bayesian modeling [15]. Another group of methods combined RPCA with training mechanisms. For instance, Yang's low-rank representation method decomposed vocals and accompaniments using pre-trained low-rank matrices [22] and Sprechmann et al. proposed a real-time method using low-rank modeling with neural networks [17]. Although these supervised learning methods demonstrated very high performance, they usually offer a weaker conception of generality.

Music instruments produce signals with various kinds of fluctuations such that they can be briefly categorized into two groups, *percussive* and *harmonic*. Signals produced by percussive instruments are more consistent along the spectral axis and by harmonic instruments are more consistent along the temporal axis with little or no fluctuations. These two categories occupy a large proportion of a spectrogram with mainly vertical and horizontal lines. To extend this sense into a more general form, the fluctuations can be viewed as a sum of sinusoid modulations along the spectral axis and the temporal axis. If a signal has nearly zero modulation along one of the two axes, its energy is smoothly distributed along that axis. Conversely, if a signal has a high frequency of modulation along one axis, then its energy becomes scattered along that axis. Therefore, if one can decipher the modulation status of a signal, one may be able to identify the instrument type of the signal. An algorithm utilizing mo-

© Frederick Yen, Yin-Jyun Luo, Tai-Shih Chi. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Frederick Yen, Yin-Jyun Luo, Tai-Shih Chi. "Singing Voice Separation using Spectro-Temporal Modulation Features", 15th International Society for Music Information Retrieval Conference, 2014.

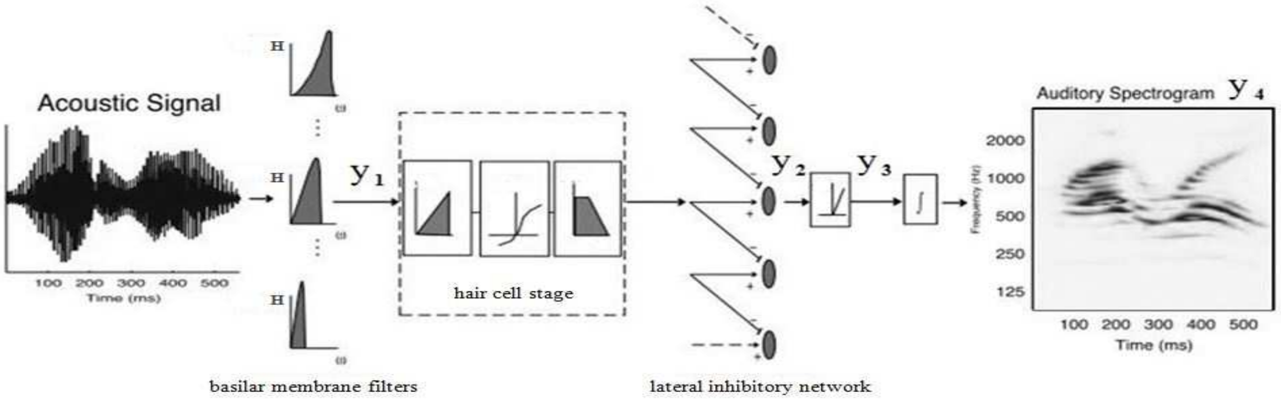

**Figure 1**. Stages of the cochlear module, adopted from [2].

dulation information can be seen in [1], where Barker et al. combined the modulation spectrogram (MS) with non-negative tensor factorization (NTF) to perform speech separation from mixtures of speech and music.

Although the above mentioned engineering approaches produce promising results, human's tremendous ability in sound streams separation makes a biomimetic approach interesting to investigate. Based on neurophysiological evidences, it is suggested that neurons of the auditory cortex (A1) respond to both spectral modulations and temporal modulations of the input sounds. Accordingly, a computational auditory model was proposed to model A1 neurons as spectro-temporal modulation filters [2]. This concept of spectro-temporal modulation decomposition has inspired many approaches in various engineering topics, such as using spectro-temporal modulation features for speaker recognition [12], robust speech recognition [18], voice activity detection [10], and sound segregation [6].

Since modulations are important for music signal categorization, this modulation-decomposition auditory model is used as a pre-processing stage for singing voice separation in this paper. Our proposed unsupervised algorithm adapts this two-stage auditory model, which decodes the spectro-temporal modulations of a T-F unit, to extract modulation based features and performs singing voice separation under the CASA framework. This paper is organized as follows. A brief review of the auditory model is presented in Section 2. Section 3 describes the proposed method. Section 4 shows evaluation and results. Lastly, Section 5 draws the conclusion.

## 2. SPECTRO-TEMPORAL AUDITORY MODEL

A neuro-physiological auditory model is used to extract the modulation features. The model consists of an early cochlear (ear) module and a central auditory cortex (A1) module.

### 2.1 Cochlear Module

As shown in Figure 1, the input sound goes through 128 overlapping asymmetric constant-Q band-pass filters ($Q_{3dB} \gg 4$ ) whose center frequencies are uniformly distributed over 5.3 octaves with the 24 filters/octave frequency resolution. These constant-Q filters mimic the frequency selectivity of the cochlea. Outputs of these filters are then transformed through a non-linear compression stage, a lateral inhibitory network (LIN), and a half-wave rectifier cascaded with a low-pass filter. The non-linear compression stage models the saturation caused by inner hair cells, the LIN models the spectral masking effect, and the following stage serves as an envelope extractor to model the temporal dynamic reduction along the auditory pathway to the midbrain. Outputs of the module from different stages are formulated below:

$$y_1(t, \omega) = s(t) *_t h(t;\omega) \tag{1}$$

$$y_2(t, \omega) = g\big(\partial_t y_1(t, \omega)\big) *_t \ell(t) \tag{2}$$

$$y_3(t, \omega) = \max\left(\partial_\omega y_2(t, \omega), 0\right) \tag{3}$$

$$y_4(t, \omega) = y_3(t, \omega) *_t \mu(t;\tau) \tag{4}$$

where $s(t)$ is the input signal; $h(t;\omega)$ is the impulse response of the cochlear filter with center frequency $\omega$ ; $*_t$ denotes convolution in time; $g(\cdot)$ is the nonlinear compression function; $\partial_t$ is the partial derivative of $t$ ; $\ell(t)$ is the membrane leakage low-pass filter; $\mu(t;\tau) = e^{-t/\tau} \cdot u(t)$ is the integration window with the time constant $\tau$ to model current leakage of the midbrain; $u(t)$ is the step function. Detailed descriptions of the cochlear module can be found in [2].

The output $y_4(t, \omega)$ of the module is the auditory spectrogram, which represents the neuron activities along time and log-frequency axis. In this work, we bypass the non-linear compression stage by assuming input sounds are properly normalized without triggering the high-volume saturation effect of the inner hair cells.

### 2.2 Cortical Module

The second module simulates the neural responses of the auditory cortex (A1). The auditory spectrogram $y_4(t, \omega)$ is analyzed by cortical neurons which are modeled by two-dimensional filters tuned to different spectro-temporal modulations. The rate parameter (in Hz) characterizes the velocity of local spectro-temporal envelope

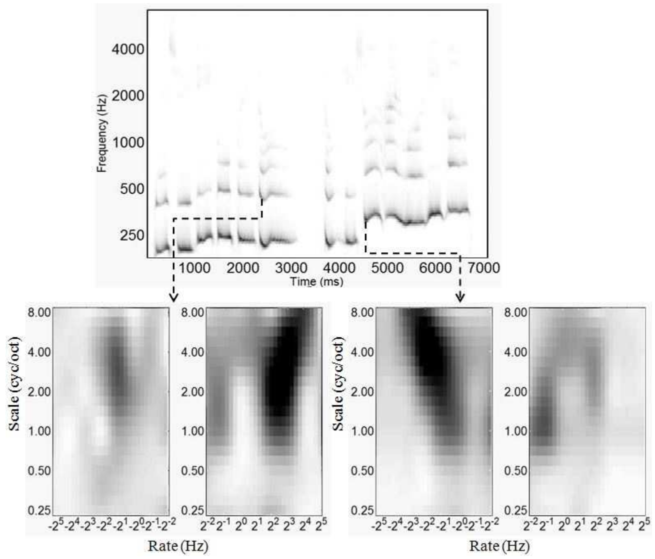

**Figure 2**. Rate-scale outputs of the cortical module to two T-F units of the auditory spectrogram of the 'Ani_2_03.wav' vocal track in MIR-1K [9].

variation along the temporal axis. The scale parameter (in cycle/octave) characterizes the density of the local spectro-temporal envelope variation along the log-frequency axis. Furthermore, the cortical neurons are found sensitive to the direction of the spectro-temporal envelope. It is characterized by the sign of the rate parameter in this model, with negative for the upward direction and positive for the downward direction.

From functional point of view, this module performs a spectro-temporal multi-resolution analysis on the input auditory spectrogram in various rate-scale combinations. Outputs of various cortical neurons to a single T-F unit of the spectrogram demonstrate the local spectro-temporal modulation contents of the unit in terms of the rate, scale and directionality parameters.

Figure 2 shows rate-scale outputs of two T-F units in an auditory spectrogram of a vocal clip. The rate-scale output is referred to as the rate-scale plot in this paper. The rate and scale indices are $\pm 2^{-2} \sim \pm 2^{5}$ and $2^{-2} \sim 2^{3}$, respectively. The strong responses of the plots correspond to the variations of singing pitch envelopes resolved by the rate and scale parameters and the moving direction of the pitch. Detailed description of the cortical module is available in [3].

## 3. PROPOSED METHOD

A schematic diagram of the proposed algorithm is shown in Figure 3. The following sections will discuss each part in details.

### 3.1 Feature Extraction

According to the spectral and temporal behaviors observed on the auditory spectrogram, components of a musical piece are characterized into three categories, *harmonic*, *percussive* and *vocal*. Harmonic components have steady energy distributions over time and have clear formant structures over frequency. Each percussive component has impulsive energy concentrated in a short period of time and has no obvious harmonic structure. Vocal components possess harmonic structure and their energy is distributed along various time periods. Interpreting the above statements from the rate-scale perspective, several general properties can be drawn. Harmonic components can be usually regarded as having low rate and high scale modulations. It means that they have relatively slow energy change along time and rapid energy change along the log-frequency axis due to the harmonic structures. In contrast, percussive components typically show quick energy change along time and energy spreading along the whole log-frequency axis, such that they possess high rate and low scale modulations. Vocal components are often recognized as a mix version of the harmonic and percussive components with characteristics sometimes considered more similar to harmonics. Different types of singing or vocal expression can result in various values of rate and scale. Figure 4 shows some examples of rate-scale plots of components from the three categories.

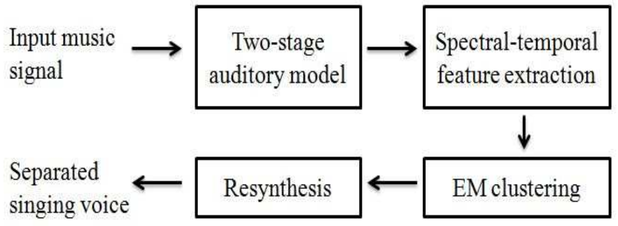

**Figure 3**. Block diagram of the proposed algorithm.

Given an auditory spectrogram $y_4 \in \mathcal{R}^{m \times n}$ transformed from an input music signal $s(t)$, the rate-scale plots of the T-F units are generated. As a pre-process, in order to prevent extracting trivial data from nearly inaudible T-F units of the auditory spectrogram, we leave out the T-F units that have energy less than 1% of the maximum energy of the whole auditory spectrogram. With the rest of the T-F units, we obtain the rate-scale plot of each unit and proceed to the feature extraction stage.

For each rate-scale plot, the total energies of the negative and positive rate side are compared. The side with greater energy is determined as the dominant plot. From the dominant plot, we extract 11 features as shown in Table 1. The features are selected by observing the rate-scale plots with some intuitive assumptions of the physical properties which distinguish between harmonic, percussive and vocal. The first 10 features are obtained by computing the energy ratio of two different areas on the rate-scale plot. For example, as shown in Table 1, the first feature is the ratio of the total modulation energy of scale = 1 to the total modulation energy of scale = 0.25. The low scales, such as 0.25 and 0.5, capture the degree of the

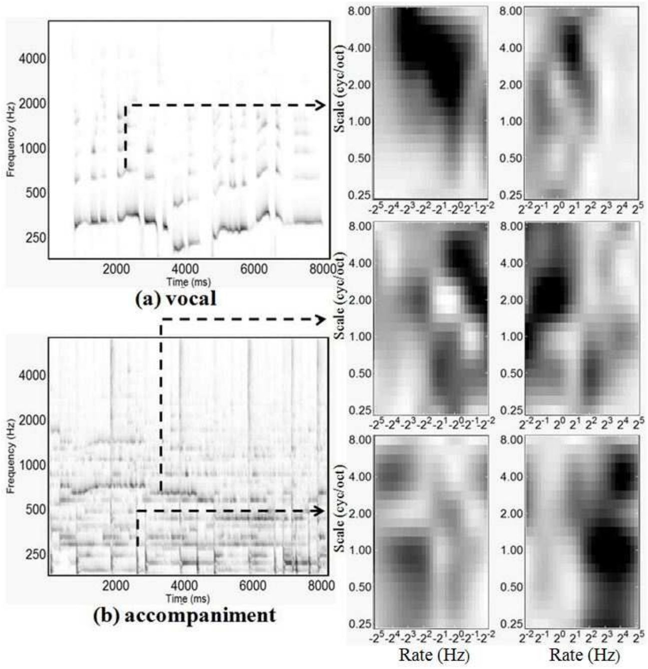

**Figure 4.** (a) Rate-scale plot from the vocal track of ‘Ani_4_07’ in MIR-1K. The modulation energy is mostly concentrated in the middle and high scales for a unit with a clear harmonic structure. (b) Rate-scale plots from the accompanying music track of ‘Ani_4_07’. The upper plot shows energy concentrating at low rates for a sustained unit. The lower plot shows energy concentrating at high rates for a transient unit.

flatness of the formant structure while the high scales, such as 1, 2, 4 and 8, capture the harmonicity with different frequency spacing between harmonics. Therefore, the first four features can be thought as descriptors which distinguish harmonic from percussive using spectral information. The fifth to the seventh features capture temporal information which can distinguish sustained units from transient units.

The feature values are saved as feature vectors and then grouped as a feature matrix F $\in \mathcal{R}^{k \times i}$ for clustering, where $k$ is the number of features and $i$ is the number of total valid units in the auditory spectrogram.

### 3.2 Unsupervised Clustering

In the unsupervised clustering stage, a spectrogram is divided into three parts and clustering is performed for each part. Based on hearing perception, the frequency resolution is higher at lower frequencies while the temporal resolution is higher at higher frequencies [14]. Due to the frequency resolution of the constant-Q cochlear filters/channels in the auditory model, the auditory spectrogram can only resolve about ten harmonics [11]. To handle different resolutions, the spectrogram is separated into three sub-spectrograms with overlapped frequency ranges. The three sub-spectrograms consist of channel 1 to channel 60, channel 46 to channel 75, and channel 61 to channel 128, respectively, with overlaps of 15 channels.

| Scale | Rate |
|---|---|
| 1 : 0.25 | all |
| 2 : 0.25 | all |
| 4 : 0.25 | all |
| 8 : 0.25 | all |
| (0.25, 2, 4) | (1, 2) : (0.25, 0.5, 1, 2, 16, 32) |
| (0.25, 2, 4) | (0.25, 0.5) : (0.25, 0.5, 1, 2, 16, 32) |
| (0.25, 2, 4) | (16, 32) : (0.25, 0.5, 1, 2, 16, 32) |
| (0.25, 0.5) : all | all |
| (1, 2) : all | all |
| (4, 8) : all | all |
| (0.25) | all |

**Table 1.** Eleven extracted modulation energy features

The clustering step is performed using the EM algorithm to group data into three unlabelled clusters. The EM algorithm assigns a probability set to each T-F unit showing its likelihood of belonging to each cluster. Note that in spectrogram representations, the sound sources are superimposed on top of each other. It implies that one T-F unit may contain energy from more than one source. Therefore, in this work, if one T-F unit has a probability set in which the second highest probability is higher than 5%, that particular T-F unit will also be labelled to the second high probability cluster. It means one unit may eventually appear in more than one cluster. The parameter 5% was empirically determined. Each of the three sub-spectrograms is clustered into three groups. Total of nine groups are generated and merged back into three whole spectrograms by comparing the correlations of the overlapped channels between different groups. Each of the three whole spectrograms represents the extracted harmonic, percussive, and vocal part of the music mixture. With no prior information about the labels of the three whole spectrograms, the effective mean rate-scale plot of each spectrogram is examined. The effective mean rate-scale plot is the mean of rate-scale plots of the T-F units with energy higher than 20% of the maximum energy in that spectrogram. The total modulation energy of rate = 1, 2 Hz and scale = 0.25, 2, 4 cycle/octave is calculated from the effective mean rate-scale plot and referred to as Ev, which is used as the criterion to select the vocal spectrogram. The one with the maximum Ev value is picked as the vocal spectrogram since Ev catches modulations related to the formant structure (scale = 0.25), the harmonic structure (scale = 2 and 4) and the singing rate (rate = 1 and 2) of singing voices.

The vocal spectrogram is then synthesized to an estimated signal using the auditory model toolbox [24]. The nonlinear operation of the envelope extractor in the cochlear module makes perfect synthesis impossible, thus causing a general result of loss of higher frequencies of the signal. Detailed computations are shown in [2].

## 4. EVALUATION RESULTS

The MIR-1K [9] is used as the evaluation dataset. It cont-

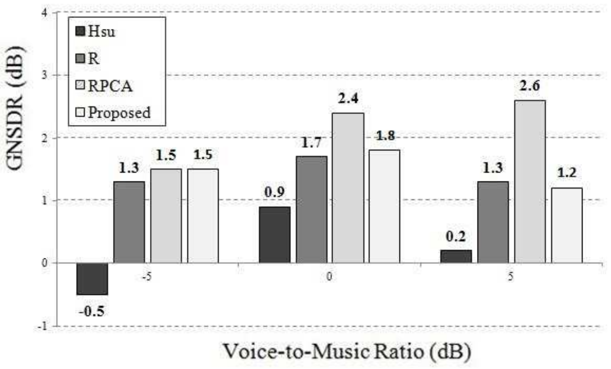

**Figure 5**. GNSDR comparison at voice-to-music ratio of -5, 0, and 5 dB with existing methods.

ains 1000 WAV files of karaoke clips sung by amateur singers. The length of each clip is around 4~13 seconds. The vocal and music accompaniment parts were recorded in the right and the left channels separately. In this experiment, we mixed two channels in -5, 0, 5 dB SNR (signal to noise ratio, i.e., vocal to music accompaniment ratio) for test. To assess the quality of separation, the source-to-distortion ratio (SDR) [21] is used as the objective measure. The ratios are computed by the BSS Eval toolbox v3.0 [23]. Following [9], we compute the normalized SDR (NSDR) and the weighted average of NSDR, the global NSDR (GNSDR), with the weighting proportional to the length of each file. To have a fair comparison, we compare our method with other unsupervised methods, which extract vocal clips only through one major stage. The compared algorithms are listed below:

I. **Hsu**: the approach proposed in [9] that performs unvoiced sound separation combined with the pitch-based inference method in [13].

II. **R** (REPET with soft masking): the approach proposed in [16] that computes a repeating background structure and extract vocal with soft time-frequency masking.

III. **RPCA:** a matrix decomposition method applying robust principal component analysis proposed by Huang et al. [8].

From Figure 5, we can observe that the proposed method has the highest performance tied with RPCA in the -5 dB SNR condition. In 0 and 5 dB SNR conditions, the performance of the proposed method is comparable to the performance of REPET.

## 5. CONCLUSION

In this paper, we propose a singing voice separation method utilizing the spectral-temporal modulations as clustering features. Based on the energy distributions on the rate-scale plots of T-F units, the vocal signal is extracted from the auditory spectrogram and the separation performance is evaluated using the MIR-1K dataset. Our proposed CASA-based masking method outperforms the CASA-based system in [9] and has comparable performance to the masking-based REPET in all SNR conditions. When compared with the subspace RPCA method, our proposed method has comparable performance only in the -5 dB SNR condition. These results demonstrate the effectiveness of the spectral-temporal modulation features for analyzing music mixtures. As this proposed method only applies a simple EM algorithm for clustering, harmonic mismatches and artificial noises are yet to be discussed.

The future work will be focused on applying more advanced classifiers for more accurate separations and adopting a two-stage mechanism like HPSS to discard percussive and harmonic components sequentially. The other potential work is to implement the proposed spectro-temporal modulation based method in the Fourier spectrogram domain [4] to mitigate synthesis errors injected by the projection-based reconstruction process of the auditory model.

## 6. ACKNOWLEDGEMENTS

This research is supported by the National Science Council, Taiwan under Grant No NSC 102-2220-E-009-049 and the Biomedical Electronics Translational Research Center, NCTU.

## 7. REFERENCES

[1] T. Barker and T. Virtanen, "Non-negative tensor factorization of modulation spectrograms for monaural sound source separation," *Proc. of Interspeech*, pp. 827-831, 2013.

[2] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, Vol. 118, No. 2, pp. 887-906, 2005.

[3] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, Vol. 106, No. 5, pp. 2719-2732, 1999.

[4] T.-S. Chi and C.-C. Hsu, "Multiband analysis and synthesis of spectro-temporal modulations of Fourier spectrogram," *J. Acoust. Soc. Am.*, Vol. 129, No. 5, pp. EL190-EL196, 2011.

[5] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation, "*IEEE J. of Selected Topics on Signal Process.*," Vol. 5, No. 6, pp. 1180-1191, 2011.

[6] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation, " *J. Acoust. Soc. Am.*, Vol. 124, No. 6, pp. 3751-3771, 2008.

[7] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorization techniques," *ISAST Trans. on Electron. and Signal Process.*, Vol. 4, No. 1, pp. 62-73 (ISSN 1797-2329), 2010.

[8] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," Porc. *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, pp. 57-60, 2012.

[9] 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. on Audio, Speech, and Language Process.*, Vol. 18, No. 2, pp. 310-319, 2010.

[10] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Voice activity detection based on frequency modulation of harmonics," *IEEE Int. Conf. on Acoust. , Speech and Signal Process.*, pp. 6679-6683, 2013.

[11] D. Klein, and S. A. Shamma, "The case of the missing pitch templates: how harmonic templates emerge in the early auditory system," *J. Acoust. Soc. Am.*, Vol. 107, No. 5, pp. 2631-2644, 2000.

[12] H. Lei, B. T. Meyer, and N. Mirghafori, "Spectro-temporal Gabor features for speaker recognition," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, pp. 4241-4244, 2012.

[13] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. on Audio, Speech, and Language Process.*, Vol. 15, No. 4, pp. 1475-1487, 2007.

[14] B. C. J. Moore: *An Introduction to the Psychology of Hearing 5$^{th}$ Ed.*, Academic Press, 2003.

[15] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, “Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs, "*IEEE Trans. on Audio, Speech, and Language Process.*," special issue on Blind Signal Proc. for Speech and Audio Applications, Vol. 15, No. 5, pp. 1564-1578, 2007.

[16] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," *IEEE Trans. on Audio, Speech, and Language Process.*, Vol. 21, No. 1, pp. 73-84, 2013.

[17] P. Sprechmann, A. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," *Proc. of the Int. Soc. for Music Inform. Retrieval Conf.*, pp. 67–72, 2012.

[18] R. M. Stern and N. Norgan, "Hearing is believing: biologically inspired methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, Vol. 29, No. 6, pp. 34–43, 2012.

[19] H. Tachibana, N. Ono, and S. Sagayama, "Singing Voice Enhancement in Monaural Music Signals Based on Two-stage Harmonic/Percussive Sound Separation on Multiple Resolution Spectrograms," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, Vol. 22, No. 1, pp. 228-237, 2014.

[20] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," *Proc. of the Int. Soc. for Music Inform. Retrieval Conf.*, pp. 337–344, 2005.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Process.*, Vol. 14, No. 4, pp. 1462-1469, 2006.

[22] Y. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," *Proc. of the Int. Soc. for Music Inform. Retrieval Conf.*, pp. 427-432, 2013.

[23] http://bass-db.gforge.inria.fr/bss_eval/

[24] http://www.isr.umd.edu/Labs/NSL/nsl.html