MUSIC/VOICE SEPARATION USING THE 2D FOURIER TRANSFORM

*Prem Seetharaman, Fatemeh Pishdadian, Bryan Pardo**

Northwestern University
Electrical Engineering and Computer Science
Evanston, IL

ABSTRACT

Audio source separation is the act of isolating sound sources in an audio scene. One application of source separation is singing voice extraction. In this work, we present a novel approach for music/voice separation that uses the 2D Fourier Transform (2DFT). Our approach leverages how periodic patterns manifest in the 2D Fourier Transform and is connected to research in biological auditory systems as well as image processing. We find that our system is very simple to describe and implement and competitive with existing unsupervised source separation approaches that leverage similar assumptions.

Index Terms— Audio source separation, singing voice extraction, 2DFT, auditory scene analysis, automatic karaoke, foreground/background separation, image processing

1. INTRODUCTION

Audio source separation is the act of isolating sound sources in an audio scene. Examples of source separation include isolating the bass line in a musical mixture, isolating a single voice in a loud crowd, and extracting the lead vocal melody from a song. Automatic separation of auditory scenes into meaningful sources (e.g. vocals, drums, accompaniment) would have many useful applications. These include melody transcription [1], audio remixing [2], karaoke [3], and instrument identification [4].

One application of source separation is singing voice extraction. A variety of approaches have been used for singing voice extraction, the vast majority of which use the spectrogram as the input representation. Examples include Non-negative matrix factorization [5], deep learning-based approaches [6], a source filter model with melodic smoothness constraints [7] and a multi-kernel framework [8].

One of the simplest and most robust approaches for singing voice extraction is to leverage repetition. REPET-SIM [9] uses repetition in the spectrogram by using the similarity matrix to find similar frames. Huang et al. [10] separate a low-rank background (the accompaniment) from a sparse foreground (the singing voice) using robust principal component analysis. The most closely related work to ours is REPET [3], which finds periodic repetition in a magnitude spectrogram, separating a periodic repeating background (accompaniment) from a non-periodic foreground (vocals). In this work we describe a novel, simple method to separate the periodic from the non-periodic audio that leverages the two dimensional Fourier transform (2DFT) of the spectrogram. The properties of the 2DFT let us separate the periodic from the non-periodic without the need to create an explicit model of the periodic audio and without the need to find the period of repetition, both of which are required in REPET.

The 2DFT has been used in music information retrieval for cover song identification [11] [12] and music segmentation [13]. There is also some prior work in audio source separation that uses the 2DFT as the input representation. Stöter et al. [14] apply the 2DFT to small 2D patches of the spectrogram. Pishdadian et al. [15] further refined this representation by using a multi-resolution 2D filter bank instead of fixed-size 2D patches. Both approaches use the 2DFT to differentiate modulation characteristics (e.g. vibrato, trills) of distinct sources and separate them from one another. These works both focus on separation of harmonic sources with the same fundamental frequencies (unisons) in very short excerpts of audio. Neither focuses on separating periodic from non-periodic patterns in long audio segments and both required the creation of a more complicated, tiled representation using the 2DFT. We present a novel singing voice extraction technique to separate periodic from non-periodic audio via a single 2DFT of the spectrogram, with no need to create a more complex multi-resolution filter bank.

2. PROPOSED METHOD

Our approach leverages the fact that musical accompaniment will typically have some amount of periodic repetition, while the vocals will be relatively aperiodic. Given this insight, we use the 2DFT to analyze the audio spectrogram and borrow a technique from image processing to perform singing voice extraction.

*This work was supported by NSF Grant 1420971.

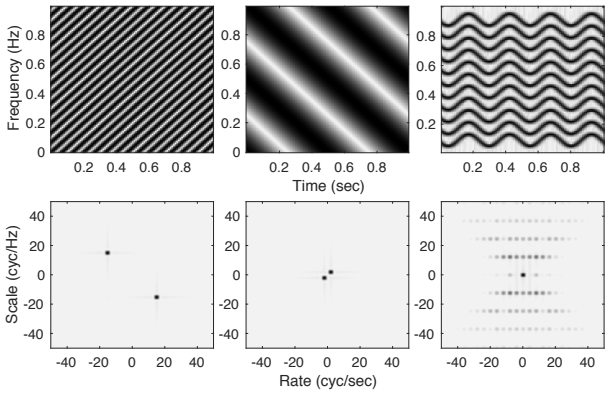

Figure 1: Examples of time-frequency-domain signals (top row) and their associated magnitude 2D Fourier transforms (bottom row). The left two show 2D sinusoids and the rightmost plot shows a more complex 2D signal. Darker colors show higher values in all plots.

2.1. The 2D Fourier Transform

The 2DFT is an essential tool for image processing, just as the 1DFT is essential to audio signal processing. The 2DFT decomposes images into a summation of weighted and phase-shifted 2D sinusoids [16]. We apply a 2DFT to the magnitude spectrogram of audio mixtures to detect and extract particular patterns such as temporal repetitions. We refer to the vertical and horizontal dimensions of the 2D transform domain as *scale* and *rate*. These terms are borrowed from studies of the auditory system in mammals [17] [18][19], which have shown that the primary auditory cortex uses representations capturing the spectro-temporal modulation patterns of audio signals. In this context, scale corresponds to the spread of spectral energy (e.g. frequency modulation depth) as well as frequency-domain repetitions (e.g. overtones) and rate corresponds to temporal repetitions (e.g. repeating percussive patterns).

In Figure 1, the left and middle columns show illustrative examples of 2D (time-frequency domain) sinusoids and their 2DFTs (scale-rate domain). A 2D sinusoid is represented by a pair of peaks in the transform domain, where the orientation of the peaks with respect to axes (upward or downward) is the opposite of the orientation of the sinusoid. The rate of repetitions across the frequency and time axes are reflected by the absolute value of scale and rate respectively. The right column shows a more complex pattern which can be decomposed into a number of 2D sinusoids using the 2DFT.

A common task in image processing is to remove noise from images. One particular denoising application is the removal of periodic noise, which can be the result of artifacts in the image capture instrument. A straightforward technique to removing periodic noise from an image is by recognizing

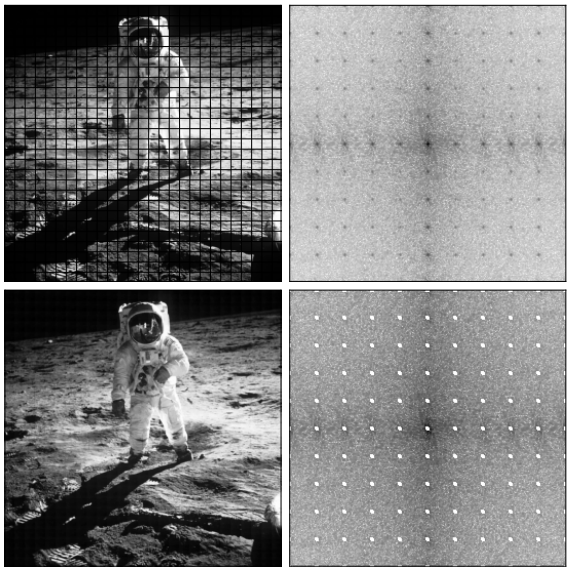

Figure 2: An example of periodic noise removal. The noisy image (upper left) is denoised by taking its 2DFT (upper right), removing local peaks that correspond to the repeating pattern (lower right) and inverting the 2DFT to obtain the denoised image (lower left).

that periodic noise will appear as a set of peaks in the 2DFT domain (see Figure 2). When 2DFT-domin peaks are masked out, one can invert the resulting representation to produce an image without the periodic noise.

In many audio signals (e.g. music), a non-periodic foreground source (e.g. a singing voice) is often accompanied by a periodic background source (e.g. a repetitive musical accompaniment). Our work adapts the idea of periodic noise removal in images to the audio realm by applying it to the magnitude spectrogram. By masking peaks in the 2DFT of the spectrogram, we can separate the periodic background from the non-periodic foreground. We now describe this algorithm for music/voice separation in more detail.

2.2. Music/voice separation

Let $x(t)$ denote a single-channel time-domain audio signal and $X(\omega, \tau)$ its complex Short-time Fourier Transform (STFT), where ω is frequency and τ is time. Our goal is to model the background music based on a repeating pattern in the magnitude plot of $X(\omega, \tau)$, also called the spectrogram. To this end, all the processing in our algorithm will be performed on $|X(\omega, \tau)|$, where $|.|$ denotes the magnitude operator. Periodically repetitive patterns in the magnitude spectrogram will appear as peaks in the 2DFT of the spectrogram, which reduces a general pattern recognition approach in the time-frequency domain into peak picking in the scale-rate domain.

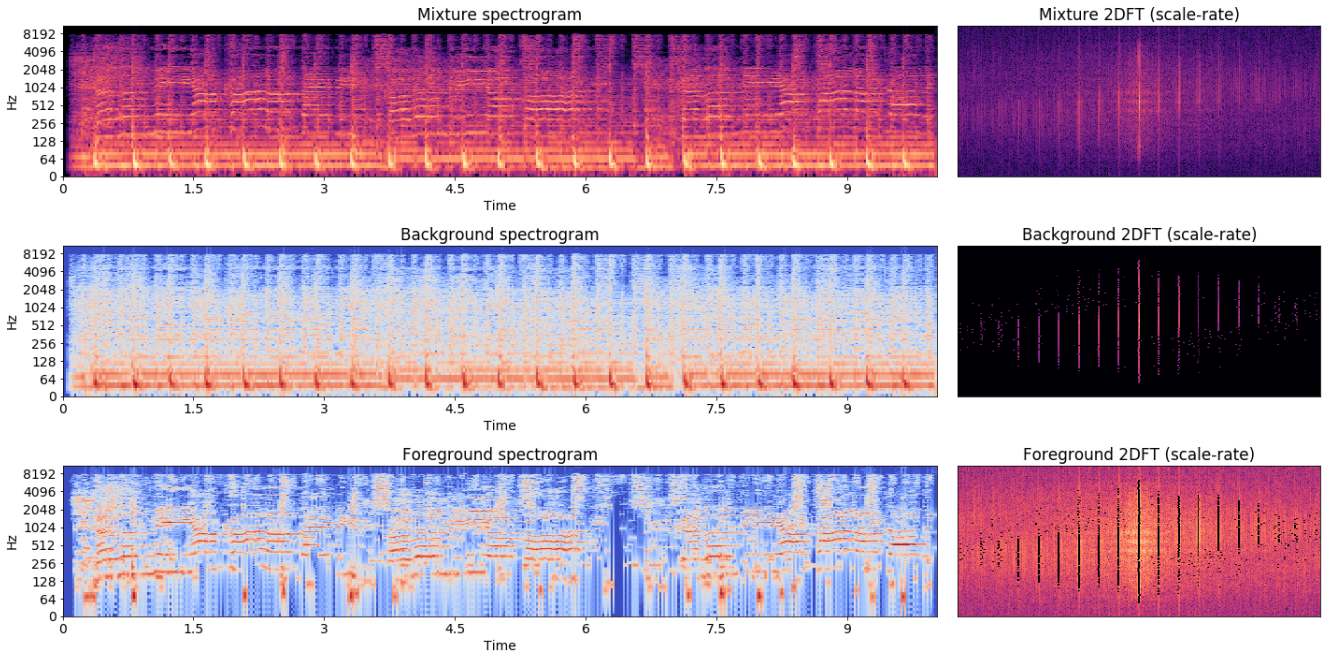

Figure 3: Separation using the 2D Fourier Transform (2DFT). In the first row, the left panel shows the mixture spectrogram and the right panel its 2DFT. In the second row, we apply our peak picking technique along the rows of the 2DFT to get a background 2DFT. Then, we invert this 2DFT and apply masking to the mixture spectrogram to get the background spectrogram. In the third row, we show everything from the rest of the 2DFT (i.e. the non-peaks), which contains the singing voice.

The scale-rate representation of the spectrogram will be denoted by $\tilde{X}(s, r)$, where s and r, stand for scale and rate respectively. The relationship between the spectrogram and its scale-rate transform can then be formulated as

$$\tilde{X}(s, r) = \mathcal{FT}_{2D}\{|X(\omega, \tau)|\}, \tag{1}$$

where $\mathcal{FT}_{2D}\{.\}$ denotes the two-dimensional Fourier transform. $\tilde{X}(s, r)$ contains complex values. The magnitude of $\tilde{X}(s, r)$ contains peaks corresponding to periodically repeating elements in the time-frequency domain. Therefore, the core of our algorithm is to locate peaks in the magnitude of the scale-rate transform (2DFT) and mask the peaks to separate the repeating accompaniment from the singing voice. We pick peaks by comparing the difference between the maximum and minimum magnitude values over a neighborhood surrounding each point in the scale-rate domain to some threshold. In this work, the threshold, denoted by γ, is set to the standard deviation of all $|\tilde{X}(s, r)|$ values.

The neighborhood for peak-picking can be of an arbitrary shape. For this work, we restrict our neighborhood shape to be a simple rectangle in the 2DFT domain. We denote the center of an arbitrary rectangular neighborhood by $c = (s_c, r_c)$, and the neighborhood surrounding this point by $N(c)$. The dimensions of the neighborhood along the scale and rate axes are tunable parameters in our algorithm.

The repeating accompaniment manifests as a series of peaks along the rate axis. Because of this, our neighborhood is shaped to find peaks along the rate axis. In this work, the size of the neighborhood along the scale axis is 1. In our experiments, we vary the size of this neighborhood along the rate axis between 15 and 100 frames in the 2DFT domain. Smaller values for the shape result in leakage from the singing voice into the accompaniment, while larger values result in leakage from accompaniment into singing voice.

Let α_c denote the range of $|\tilde{X}(s, r)|$ values over the neighborhood, that is

$$\alpha_c = \max_{N(c)} |\tilde{X}(s, r)| - \min_{N(c)} |\tilde{X}(s, r)|. \tag{2}$$

The value of the peak-picking mask, which we will refer to as the scale-rate domain *background mask* can thus be computed at c as follows

$$M_{bg}(s_c, r_c) = \begin{cases} 1 & \alpha_c > \gamma, |\tilde{X}(s_c, r_c)| = \max\limits_{N(c)} |\tilde{X}(s, r)| \\ 0 & otherwise \end{cases} \tag{3}$$

Intuitively, this is simply a way to discover local maxima in $|\tilde{X}(s, r)|$ that are above a threshold γ. It should be noted that neighborhood selection and mask value computation is performed for every single point in the scale-rate domain. We denote the computed background mask over the

Method	Voice SDR	Voice SIR	Voice SAR	Music SDR	Music SIR	Music SAR
RPCA	2.3 ± 1.5	11.0 ± 4.5	$\mathbf{2.9 \pm 4.0}$	5.0 ± 2.3	7.7 ± 2.9	10.5 ± 6.6
REPET-SIM	2.1 ± 1.5	15.2 ± 4.2	$\mathbf{2.9 \pm 3.8}$	$\mathbf{6.3 \pm 2.7}$	$\mathbf{12.5 \pm 2.7}$	$\mathbf{10.5 \pm 6.7}$
REPET	2.2 ± 1.5	$\mathbf{15.6 \pm 4.9}$	2.8 ± 3.8	5.0 ± 2.6	10.2 ± 2.7	10.4 ± 6.6
2DFT (1, 15)	2.6 ± 1.5	11.8 ± 3.9	2.8 ± 4.0	5.7 ± 2.5*	8.7 ± 3.0*	10.4 ± 6.7
2DFT (1, 35)	$\mathbf{2.7 \pm 1.6}$*	13.2 ± 3.9	2.8 ± 4.0*	5.1 ± 2.5	7.6 ± 2.9	10.4 ± 6.6*
2DFT (1, 100)	2.6 ± 1.5	13.5 ± 4.0*	2.7 ± 4.0	4.4 ± 2.4	6.7 ± 2.8	10.3 ± 6.6
Ideal Binary Mask	9.2 ± 2.7	30.0 ± 4.1	9.5 ± 3.2	14.9 ± 6.5	27.9 ± 8.0	15.2 ± 6.5

Table 1: SDR/SIR/SAR for the singing voice and the music accompaniment as extracted from the mixture. In the rows labeled 2DFT, the neighborhood shape in which we do the peak picking is shown in the parenthesis (e.g. (M, N) is M rows by N columns.) The best performance for our system is indicated by an asterisk, while the best performance across all algorithms is indicated by boldface. Note that SDR for foreground and background sources for our optimal settings are higher than those of REPET, but lower than REPET-SIM, which has the advantage of exploiting non-periodic patterns as well as periodic ones.

entire scale-rate domain representation by $M_{bg}(s, r)$. The scale-rate domain *foreground* mask can then be computed as $M_{fg}(s, r) = 1 - M_{bg}(s, r)$.

Next, we compute the separated magnitude spectrogram of the background (repeating) source from the masked version of the complex scale-rate domain representation, by taking the inverse 2DFT of the masked signal:

$$|X_{bg}(\omega, \tau)| = \mathcal{IFT}_{2D}\{M_{bg}(s, r) \odot \tilde{X}(s, r)\}, \quad (4)$$

with $\mathcal{IFT}_{2D}\{.\}$ denoting the inverse 2D Fourier transform and $\odot$ denoting element-by-element multiplication, respectively. The foreground magnitude spectrogram can be similarly computed using the foreground mask.

The separated audio is obtained by masking in the time-frequency domain. The time-frequency masks are simply computed by comparing the inverted magnitude spectrograms from the 2DFT for foreground and background:

$$M_{bg}(\omega, \tau) = \begin{cases} 1 & |X_{bg}(\omega, \tau)| > |X_{fg}(\omega, \tau)| \\ 0 & otherwise, \end{cases} \quad (5)$$

and the foreground mask as $M_{fg}(\omega, \tau) = 1 - M_{bg}(\omega, \tau)$.

In the last step, the time-domain background and foreground audio signals are recovered from the masked STFT. In short, $x_{bg}(t) = ISTFT\{M_{bg}(\omega, \tau) \odot X(\omega, \tau)\}$, where $ISTFT\{.\}$ is the Inverse Short-Time Fourier Transform, computed through the overlap-and-add method. The foreground audio signal (the singing voice) can be similarly computed by applying the foreground mask to the complex spectrogram and taking the inverse STFT. The separation process can be seen in Figure 3.

3. EVALUATION

We evaluate our approach using DSD100 [20], a dataset consisting of 100 multitrack recordings of four sources - vocals, drums, bass, and other. We label the combination of the latter three sources the accompaniment. Our task is to separate the vocals from the accompaniment. We extract 30 second clips from each multitrack example. The four sources (vocals, drums, bass, other) are combined into a mono mixture for separation. We compare our method to other methods for singing voice extraction that use an assumption of a low rank accompaniment source. These are REPET [3], REPET-SIM [21], and RPCA [10]. For our proposed method, we vary the size of the neighborhood for peak picking, as described in Section 2. We also compare to the ground truth sources. Separation performance is evaluated using the BSS Evaluation metrics [22] source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifact ratio (SAR).

SDR/SIR/SAR results are shown in Table 1 for our proposed method and competing methods. Our proposed method shows very competitive results to a variety of algorithms for source separation based on repetition. The most direct comparison is with REPET, which also performs music/voice separation via repeating pattern extraction. REPET depends on computing a precise length of the periodic pattern. If the computed length is off by even one frame, the performance of the separation will be sub-optimal. Our approach does not require estimation of the length of the period or explicit modeling of the repeating pattern. This approach connects an image processing technique to source separation. The periodic noise removal in Figure 2 and the repeating background extraction in Figure 3 are done using the same algorithm described in Section 2.[1]

4. CONCLUSION

We presented a simple and novel approach for music/voice separation. Our approach leverages how periodic patterns manifest in the scale-rate domain and is connected to research in biological auditory systems as well as image processing. We find that our system is competitive with existing unsupervised source separation approaches that leverage similar assumptions.

[1] Audio examples at https://interactiveaudiolab.github.io/demos/2dft.

5. REFERENCES

[1] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic music transcription and audio source separation," *Cybernetics &Systems*, vol. 33, no. 6, pp. 603–627, 2002.

[2] J. F. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation.," in *ISMIR*, pp. 314–319, 2006.

[3] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 73–84, 2013.

[4] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation.," in *ISMIR*, pp. 327–332, 2009.

[5] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Latent variable analysis and signal separation*, pp. 140–148, Springer, 2010.

[6] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 2135–2139, IEEE, 2015.

[7] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[8] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *Signal Processing, IEEE Transactions on*, vol. 62, no. 16, pp. 4298–4310, 2014.

[9] Z. Rafii and B. Pardo, "Online repet-sim for real-time speech enhancement," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 848–852, IEEE, 2013.

[10] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 57–60, IEEE, 2012.

[11] P. Seetharaman and Z. Rafii, "Cover song identification with 2d fourier transform sequences," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017*, 2017.

[12] T. Bertin-Mahieux and D. P. Ellis, "Large-scale cover song recognition using the 2D Fourier transform magnitude," in *International Society for Music Information Retrieval Conference*, 2012.

[13] O. Nieto and J. P. Bello, "Music segment similarity using 2d-fourier magnitude coefficients," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 664–668, IEEE, 2014.

[14] F.-R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 126–130, IEEE, 2016.

[15] F. Pishdadian, B. Pardo, and A. Liutkus, "A multiresolution approach to common fate-based audio separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 2017.

[16] J. C. Russ and R. P. Woods, "The image processing handbook," 1995.

[17] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[18] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[19] P. Ru and S. A. Shamma, "Representation of musical timbre in the auditory cortex," *Journal of New Music Research*, vol. 26, no. 2, pp. 154–169, 1997.

[20] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, vol. 9237 of *Latent Variable Analysis and Signal Separation*, (Liberec, France), pp. 387–395, Aug. 2015.

[21] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix.," in *ISMIR*, pp. 583–588, 2012.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.