An Informed Source Separation System for Speech Signals

Shuhua Zhang, Laurent Girin

GIPSA-lab, Grenoble Institute of Technology, Grenoble, France

{shuhua.zhang, laurent.girin}@gipsa-lab.grenoble-inp.fr

Abstract

In two previous papers, we proposed an audio Informed Source Separation (ISS) system which can achieve the separation of $I > 2$ musical sources from linear instantaneous stationary stereo (2-channel) mixtures, based on audio signal's natural sparsity, pre-mix source signals analysis, and side-information embedding (within the mix signal). In the present paper and for the first time, we apply this system to mixtures of (up to seven) simultaneous speech signals. Compared to the reference MPEG-4 Spatial Audio Object Coding system, our system provides much cleaner separated speech signals (consistently 10–20 dB higher Signal to Interference Ratios), revealing strong potential for audio conference applications.

Index Terms: underdetermined source separation, speech mixture, speech signals sparsity, signal compression

1. Introduction

Source separation aims at recovering I unobserved source signals $s_i[n]$, $i \in [1, I]$, from J observations of their mixture $x_j[n]$, $j \in [1, J]$. In this paper, we consider speech signals, and we address the stereo *underdetermined* configuration, where $I > 2$ speech signals have to be separated from only $J = 2$ channels. This is a difficult configuration, that cannot be processed by Blind Source Separation (BSS) / Independent Components Analysis (ICA) methods developed for (over)determined mixtures ($J \geq I$) [1, 2], and that is better addressed with techniques based on sparse Time-Frequency (TF) representations of audio signals [3, 4, 5], or other a priori information on the source signals and the mixture process.

In [6] we proposed an Informed Source Separation (ISS) system, based on audio signals sparsity and a two-step coder-decoder structure, and dedicated to music demixing. The coder corresponds to the music signal production level (e.g., music recording/mixing in studio) where the source signals are assumed to be available and the mixing process is controlled. The decoder corresponds to the personal music player, where only the stereo mix signal is available. Parameters that characterize the source signals and the mixing process are embedded into the mixture signal at the coder level, so that they can be retrieved at the decoder and exploited for source signals separation from the mix signal. For instance, the mixture process is Linear Instantaneous Stationary Stereo (LISS) (aka constant-gain stereo panning) and the side-information consists of i) the mixture matrix, and ii) the indexes of the two predominant sources in each TF bin as provided by "Oracle" estimation at the coder. The decoder uses those information to perform local mixture inversion in each TF bin. The separation performances obtained by this system are quite impressive: Signal to Distortion Ratio (SDR) [7] gains about 20 dB are obtained for all sources of 5-source music mixtures, enabling realistic applications such as generalized karaoke/soloing.

Such ISS system also has a strong potential for audio conference applications, to separate and respatialize different simultaneous speakers. Therefore, in the present paper, and for the first time, we apply the ISS system to speech signals, and assess the separation performances for "difficult" mixtures of continuous speech with many speakers (up to 7). Also, we compare our system with the MPEG-4 Spatial Audio Object Coding (SAOC) system [8], which is also a system dedicated to restitute audio sources ("audio objects") from downmix signals and side-information (composed of inter-object spatial cues). The technical positioning of our system with respect to SAOC is briefly discussed in Section 3. Note that SAOC applications also concern both music and speech demixing/remixing.

This paper is organized as follows. The proposed system is described in Section 2. Results obtained for speech mixtures as well as comparison with SAOC are presented in Section 3, and Section 4 concludes the paper.

2. The ISS system

The ISS system used in this work is depicted in Fig. 1. It is similar to the one presented in [6]. However, the parameters of the ISS process are adapted to speech signals. Also, in [6], uncompressed (16-bit PCM) mixture signals were considered (and the side-information was embedded into the mix signal using a high-capacity watermarking technique [9]). In the present framework, the speech mixture signal generally has to be transmitted in dedicated channels, hence we consider the use of both uncompressed (16-bit PCM) and compressed mixture signals (MPEG-2/4 AAC [10]). The side-information is assumed to be either embedded into the bitstream or transmitted on a dedicated channel.

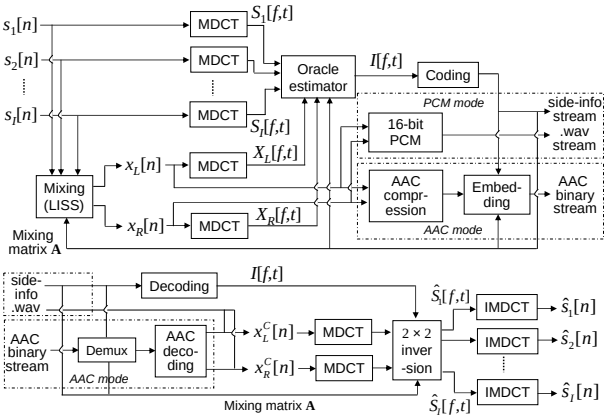

Figure 1: Diagram of the ISS system.

2.1. Time-frequency decomposition

The speech source signals are non-stationary, with quite large temporal and spectral variability, and they possibly strongly overlap in the time domain. Using TF representation is known to exhibit their natural sparsity, i.e., much lower overlapping of signals in the TF domain, a foundation for "sparse separation" methods [4, 5]. A linear invertible transform is required, so that the source separation problem remains linear/instantaneous in the transformed domain. And preferably, the transform sparsifies signals maximally and suppresses the blocking artifact intrinsically. In this regard, we select the Modified Discrete Cosine Transform (MDCT [11]), a lapped orthogonal transform widely used in audio compression and processing. This transformation is applied on each source signal at the coder, and on each mixture channel at both coder and decoder, while the corresponding inverse transform (IMDCT) is used at the decoder to regenerate estimated time-domain source signals from separated MDCT coefficients (Fig. 1). We use here time frames of W=768 samples (approximately 48 ms for 16 kHz speech signals), with a 50% overlap between consecutive frames. This setting enables to follow the time-dynamics of speech signals while providing a frequency resolution of about 21 Hz suitable for the separation.

2.2. Local inversion of the mixture and sources selection

As in the blind method of [3] and in our previous works [6], the estimation of source signals is processed by a local inversion of the mixture signal. "Local" means that the process is considered for each TF bin $[f,t]$, and at this level, only at most $J=2$ sources are assumed to be contributing significantly to the mixture signal. Therefore, in the MDCT domain, the mixture $\mathbf{X}[f,t]=\mathbf{A}\cdot\mathbf{S}[f,t]$ at each TF bin is assumed to be locally reduced to:

$$\mathbf{X}[f,t] \approx \mathbf{A}_{\mathcal{I}_{ft}} \cdot \mathbf{S}_{\mathcal{I}_{ft}}[f,t], \tag{1}$$

where $\mathcal{I}_{ft}$ denotes the set of $J=2$ most relevant sources at TF bin $[f,t]$, i.e., the two source signals that locally "better explain" the mixture. $\mathbf{A}_{\mathcal{I}_{ft}}$ represents the 2×2 mixing sub-matrix made with the $\mathbf{A}_i$ columns of $\mathbf{A}$, $i\in\mathcal{I}_{ft}$ (which is assumed to be invertible and well-conditioned). If $\bar{\mathcal{I}}_{ft}$ denotes the complementary set of non-active (or at least poorly active) sources at TF bin $[f,t]$, the source signals are estimated by:

$$\begin{cases} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}[f,t] &= \mathbf{A}^{-1}_{\mathcal{I}_{ft}} \cdot \mathbf{X}[f,t] \\ \hat{\mathbf{S}}_{\bar{\mathcal{I}}_{ft}}[f,t] &= 0 \end{cases}. \tag{2}$$

The side-information that is transmitted between ISS coder and decoder (in addition to the mix signal) consists of i) the coefficients of the mixing matrix $\mathbf{A}$, and ii) the optimal combination of source indexes $\mathcal{I}_{ft}$ for each TF bin. This contrasts with (semi-)blind separation methods where this information have to be estimated from the mix signal only, generally in two steps which can both be a challenging task and source of significant errors.

As for the mixing matrix, the number of coefficients to be transmitted is quite low in the present LISS configuration (only I fixed coefficients for each mixing configuration, if $\mathbf{A}$ is made of normalized column vectors). Therefore, the transmission cost of $\mathbf{A}$ is negligible compared to the transmission cost of $\mathcal{I}_{ft}$, and in the following we do not detail the encoding and transmission of $\mathbf{A}$.

As for the source indexes, in the specific ISS framework, the optimal $\mathcal{I}_{ft}$ is estimated using the source signals, the matrix $\mathbf{A}$, and the mixture signals. This is done using an Oracle estimator, as introduced in [12] for providing upper bounds for the performances of source separation algorithms. Exploiting the reconstruction properties of the MDCT, the overall best separation in the time domain in the mean squared error (MSE) sense is obtained by finding the optimal combination of source signals at each TF bin separately [12]:

$$\tilde{\mathcal{I}}_{ft} = \arg\min_{\mathcal{I}_{ft}\in\mathcal{P}} \sum_{i=1}^{I} \left(\hat{S}_i[f,t] - S_i[f,t]\right)^2, \tag{3}$$

where $\mathcal{P}$ represents the set of all possible combinations $\mathcal{I}_{ft}$ and the I estimated source signals $\hat{S}_i(f,t)$ are provided by (2). If I is limited to a reasonable number of sources, $\tilde{\mathcal{I}}_{ft}$ can be found by exhaustive search, and coded with a very limited number of bits before being embedded into the mixture signal stream (see Section 2.4).

2.3. Separation from transmitted mixtures

The separation at the decoder consists of applying (2) using the MDCT coefficients $\mathbf{X}^C[f,t]$ of the transmitted mix signal $\mathbf{x}^C$, instead of the coefficients of the original mix signal $\mathbf{X}[f,t]$:

$$\begin{cases} \hat{\mathbf{S}}_{\tilde{\mathcal{I}}_{ft}}[f,t] &= \mathbf{A}^{-1}_{\tilde{\mathcal{I}}_{ft}} \cdot \mathbf{X}^C[f,t] \\ \hat{\mathbf{S}}_{\bar{\tilde{\mathcal{I}}}_{ft}}[f,t] &= 0 \end{cases}. \tag{4}$$

In the current system (Fig. 1), the mix signal at the output of the coder is either PCM or AAC bitstreams. In our previous studies, 16-bit PCM conversion has been shown to have negligible effects on separation performances. In contrast, audio compression is expected to perturb the MDCT coefficients of the mix signal significantly, depending on bitrate, and is thus likely to impair the separation performance. Also, the sparsity assumption may not hold perfectly, in which case, non-predominant but active sources will interfere as noises in the local inversion process. In Section 3, we report separation results for speech signals mixed with the LISS configuration and compressed with different settings.

2.4. Side-information coding and embedding

Instead of embedding the side-information ($\mathbf{A}$ and $\tilde{\mathcal{I}}_{ft}$) into the mix signal waveform through a high-capacity watermarking technique [6, 9], in the current system we transmit the side-information using metadata segments of the compressed binary stream. It can also be transmitted using a dedicated channel. This is similar to the spirit of SAOC [8] (see Section 3).

Since $\mathbf{A}$ is fixed and only needs to be transmitted once for the entire signal, we consider only $\tilde{\mathcal{I}}_{ft}$ here. Suppose that I is fixed for all frames and frequencies for a given mix signal, with basic entropy-coding, the side-information occupies approximately $2f_{\max}/F_s \log_2(\mathcal{I}_{ft})$ bits per sample, where $f_{\max}$ is the maximal frequency processed, and F_s is the sampling frequency of the mix signal. For example, for a 5-source mixture, we have 10 combinations of 2 active sources out of 5, leading to $\log_2 10 = 3.3$ bit per sample, or 53 kbps for the side-information, given $f_{\max}=8$ kHz and $F_s=16$ kHz.

This basic coding scheme can be refined by exploiting the fact that musical/speech sources also generally have some temporal sparsity. Before mixing, each source signal is then labeled into non-silent/silent sections, and for each MDCT frame, a I-bit code c_1 is transmitted to provide the combination of non-silent sources. For each frequency bin f, the estimation of $\tilde{\mathcal{I}}_{ft}$

by (3) is then carried out only among the non-silent sources, leading to a lower bitrate. For example, if only 4 sources out of 5 are non-silent on a given frame, $\tilde{\mathcal{I}}_{ft}$ only concerns 2 sources out of 4, then maximally $\log_2(6) = 2.6$ bits per sample. In addition, only c_1 has to be transmitted if the number of non-silent sources is lower or equal to 2. Such refinement has been shown to save up to 40% of bitrate depending on the number of sources and their degree of temporal sparsity. Finally, it appears that the distribution of $\tilde{\mathcal{I}}_{ft}$ in the TF plane is highly structured, and many lossless coding strategies can be applied to the side-information.

In the present paper, we do not detail the format of the metadata segment in the compressed bitstream, and the corresponding embedding strategy. We, however, provide in Section 3 the typical ISS side-information bitrates obtained for our test signals (and we compare them with the SAOC side-information bitrates for the same test signals).

3. Experiments on speech mixtures

3.1. Experimental settings

The ISS system has been applied to mixtures of $I = 3$, 5, and 7 simultaneous talks, downmixed to stereo (2-channel) signals. All test speech signals are 16-kHz signals taken from the TIMIT database [13], concatenated and level-adjusted to form continuous (non-stop) 5s-clips with the same volume. In the mixed stereo signal, each source signal is panned to a different azimuth, governed by a constant mixing matrix $\mathbf{A}$, to simulate speakers from different angles. To ensure good conditioning of each sub-matrix of $\mathbf{A}$ and spatial (perceived) separation of the speakers, we have

$$\mathbf{A} = \begin{bmatrix} \cos\theta_1 & \cos\theta_2 & \cdots & \cos\theta_I \\ \sin\theta_1 & \sin\theta_2 & \cdots & \sin\theta_I \end{bmatrix}, \qquad (5)$$

where $\theta_k = (2k-1)\pi/(4I)$. In such mix of dense simultaneous talks of the same volumes, we observed that in case of $I = 3$, normal people could barely understand who was saying what; in case of $I = 7$, the mix is very messy and close to cafeteria noises. So our experiments presented hard cases for the separation task, even for human beings.

We evaluated the ISS system in terms of Signal to Interference Ratios (SIRs) and SDRs as defined in [7], with and without encoding the mixture signals by AAC. When encoding mixture stereo signals, we used the AAC codecs from Nero [1]. and bitrates were set to 16, 32, and 64 kbps employing the AAC Low Complexity (LC) profile [10] (this was to avoid spectral distortion resulting from parametric coding enabled at lower bitrates).

For comparison, we also evaluated the SIRs and SDRs of SAOC using the same source and mix signals. As mentioned in the introduction, SAOC is a system dedicated to restitute "audio objects" from downmix signals and side-information. SAOC is built on the former MPEG Surround or Spatial Audio coding (SAC) system [14], which was initially designed for the respatialization of multichannel audio from downmix signals using interchannel spatial cues (transmitted as low-bitrate side-information). SAOC extends the respatialization concept of SAC to audio objects (e.g., musical instruments or individual speech signals), thus leading de facto to source separation through respatialization. However, because it was initially designed for spatialization, the SAOC decoder may not ensure sufficient separation quality for any kinds of audio scenes. In contrast, even if the overall spirit is close, our ISS system is designed specifically for source separation (of LISS mixtures) and uses specific parameters. Note that for those comparative experiments, we used the SAOC reference system provided on the MPEG website [2], which may not be an optimized version. Note also that downmixing by matrices defined in (5) fits well the model of SAOC — sources are spatialized significantly and stably — a sweet point for SAOC.

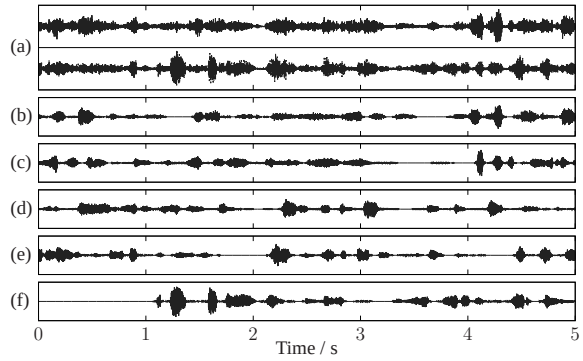

Figure 2: Waveforms of the stereo mix (a), and the $I = 5$ separated signals (b)–(f) (ISS system, AAC 32 kbps).

3.2. Separation results

As an example, the stereo mix signals and speech signals separated by the ISS system are shown in Fig. 2 for $I = 5$ and bitrate = 32 kbps. They come from male and female American speakers with different accents.

The SIRs and SDRs obtained on our test signals are provided in Fig. 3 and Fig. 4 respectively. In Fig. 3, we find that mean (across sources) SIRs obtained with the ISS system are almost all higher than or close to 30 dB, even for the extreme case of 7 simultaneous speakers. This indicates that the sparsity assumption is well hold for multiple simultaneous speech signals. The SIRs are remarkably robust to compression: differences from PCM to AAC-64 kbps and then to AAC-32 kbps are generally small (however, a large decrease to 16 kbps, of about 10 dB). This implies that mix signal compression at relatively low bitrate, say 32 kbps, is sufficient to ensure good separation of sources. The ISS system outperforms the SAOC system by 10–25 dB in SIRs. This is also clearly verified by informal listening tests: cross-talks are much more apparent in SAOC. It should also be noted that the SIRs of SAOC are almost the same for PCM and AAC at different bitrates, but depend on the number of speakers, suggesting that SIRs in SAOC are more limited by the SAOC spatialization model than by AAC encoding distortion.

Compared to the SIRs, SDRs of both the ISS system and SAOC in Fig. 4 are 5–30 dB lower. Therefore, processing artifacts dominate the overall distortions. These artifacts come from both AAC encoding and source separation. And generally, the more numerous the speakers, the more dominant the artifacts from source separation, the less important the artifacts from compression. This is shown as SDRs of the ISS system drop about 7 dB from PCM to AAC-64 kbps for $I = 3$, but drop only about 2 and less than 1 dB for $I = 5$ and 7, respectively. For $I = 7$, AAC at 64 kbps is a good trade-off between source separation quality and mix bitrate, compared to PCM.

[1] http://www.nero.com/eng/technologies-aac-codec.html

[2] http://www.itscj.ipsj.or.jp/sc29/29w42911.htm

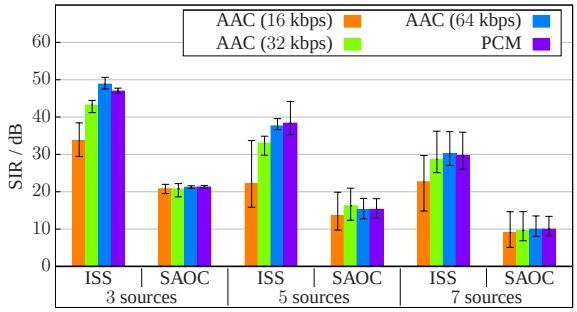

Figure 3: Mean (across sources) SIRs of the ISS system and SAOC, with and without encoding mixtures by AAC. The horizontal bars are for minimum and maximum SIRs.

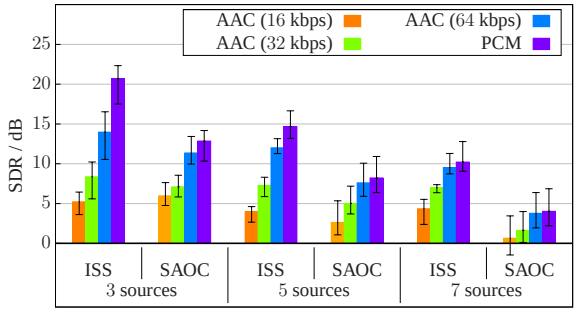

Figure 4: Mean (across sources) SDRs of the ISS system and SAOC, with and without encoding mixtures by AAC. The horizontal bars are for minimum and maximum SDRs.

Informal listening tests reveal that except for the case of $I = 7$ and bitrate $= 16$ kbps, speech signals separated by the ISS system are well intelligible (of course, they are of better quality when going toward smaller I and higher bitrates). For SAOC, intelligibility is ensured for all separated sources only when the mixtures are compressed at bitrate no less than 32 kbps (or not compressed) and the number of speakers is no more than 5. Demo sequences can be downloaded from `http://www.gipsa-lab.inpg.fr/~laurent.girin/demo/ISS-SAOC-speech.zip`.

With basic entropy coding, the side-information bitrates of the ISS system are $15.6, 40.0, 56.9$ kbps for $I = 3, 5$, and 7, respectively. These bitrates can be significantly reduced by taking the advantages of the structure of the indexes and redundancy among them. The side-information bitrates of the SAOC system are about $3/4$ lower, thanks to its advanced entropy coding scheme.

4. Conclusions

We have extended the ISS system proposed in [6] for the informed separation of uncompressed music signals to the case of uncompressed or compressed speech signals. Experiments with the extended system applied to continuous simultaneous talks of up to 7 speakers has demonstrated that local sparsity, the core assumption of the system, withstands dense mixtures of many speech signals that are impossible to distinguish for normal listeners. At different compression levels and with different numbers of speakers, the extended system is capable to separate stereo mix signals into intelligible individual speech signals, and maintains SIRs higher than 30 dB (except for one condition) while SDRs are dominated by encoding and separation artifacts. Therefore, reducing spectral artifacts due to the ISS processing is key to improving its performance. Also we will refine the entropy coding scheme of the indexes to reduce side-information bitrate and integrate the ISS system more closely into AAC to reduce computational complexity.

5. Acknowledgements

This project is supported by the French National Research Agency (ANR) - Grant CONTINT 2009 CORD 006.

6. References

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation - Independent Component Analysis and Applications*. Academic Press, 2010.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, Eds., *Independent Component Analysis*. Wiley and Sons, 2001.

[3] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," *Signal Proc.*, vol. 81, no. 11, pp. 2353–2362, 2001.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Proc.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[6] M. Parvaix and L. Girin, "Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding," in *IEEE Int. Conf. Acoust., Speech, Signal Proc.*, Dallas, Texas, 2010.

[7] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[8] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers *et al.*, "Spatial audio object coding (SAOC): The upcoming MPEG standard on parametric object based audio coding," in *124th AES Convention*, Amsterdam, The Netherlands, May 2008.

[9] J. Pinel, L. Girin, C. Baras, and M. Parvaix, "A high-capacity watermarking technique for audio signals based on MDCT-domain quantization," in *Int. Congress on Acoustics*, Sydney, Australia, 2010.

[10] ISO/IEC JTC1/SC29/WG11 MPEG, "Information technology—Generic coding of moving pictures and associated audio information—Part 7: Advanced Audio Coding (AAC)," IS13818-7(E), 2004.

[11] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 64, no. 5, pp. 1153–1161, 1986.

[12] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Proc.*, vol. 87, no. 8, pp. 1933–1950, 2007.

[13] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, USA, 1993.

[14] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. Chong, "MPEG surround—the ISO/MPEG standard for efficient and compatible multichannel audio coding," *Journal of Audio Engineering Society*, vol. 56, no. 11, pp. 932–955, Nov. 2008.