An Image-Inspired Audio Sharpness Index

Gaël Mahé, Lionel Moisan, Mihai Mitrea

An Image-Inspired Audio Sharpness Index

Gaël Mahé
LIPADE
Université Paris Descartes
Paris, France
Email: gael.mahe@parisdescartes.fr

Lionel Moisan and Mihai Mitrea
MAP5, CNRS UMR 8145
Université Paris Descartes
Paris, France
Emails: lionel.moisan@parisdescartes.fr, mihai.mitrea@telecom-sudparis.eu

***Abstract*—We propose a new non-intrusive (reference-free) objective measure of speech intelligibility that is inspired from previous works on image sharpness. We define the audio Sharpness Index (aSI) as the sensitivity of the spectrogram sparsity to the convolution of the signal with a white noise, and we calculate a closed-form formula of the aSI. Experiments with various speakers, noise and reverberation conditions show a high correlation between the aSI and the well-established Speech Transmission Index (STI), which is intrusive (full-reference). Additionally, the aSI can be used as an intelligibility or clarity criterion to drive sound enhancement algorithms. Experimental results on stereo mixtures of two sounds show that blind source separation based on aSI maximization performs well for speech and for music.**

I. Introduction

Measuring speech intelligibility is of interest in many applications where the transmission channel between the speech source and the listener can be impaired: telecommunications, public announcement, room acoustics or hearing impairment. Subjective methods based on listening tests (see [1] for a review) provide the most reliable measure of intelligibility, but they are time- and money-consuming. This is why many objective measures have been also proposed, aiming at predicting the subjective intelligibility from signal parameters alone. Another interesting property of objective measures is that they can be used as an optimization criterion in speech enhancement algorithms (e.g., for noise reduction and dereverberation).

Among the standardized methods, the Speech Intelligibility Index (SII [2]) is a weighted sum of band-specific signal to noise ratios, and the Speech Transmission Index (STI [3]) measures how the acoustic channel reduces the modulation index for various frequency bands and modulation frequencies. Several recent measures estimate in each frequency band the correlation (e.g. Short-Time Objective Intelligibility, STOI [4]) or the mutual information (SIMI [5]) between the clean and the distorted signals.

All these methods are intrusive or full-reference methods, which means that they require the knowledge of the clean signal to measure the intelligibility of the distorted signal. However, in many cases, the clean signal is not available (e.g., audio record or voice received in telephony), so that non-intrusive (or reference-free) measures are required. Most of them, like NIRA [6] or NISI [7], are based on machine-learning techniques and derive indicators from a large set of signal parameters by maximizing the correlation with reference indicators on a training corpus. The drawback of this approach is that the indicators depend on the training conditions and that they are blind to the physical grounds of intelligibility.

Another approach was proposed in [8], based on the modulation spectrogram, that is, the spectrum of the temporal envelope computed in each frequency band. The principle is that the modulation energy is concentrated around 4 Hz for clean signals and tends to spread towards high modulation frequencies in case of reverberation. Hence, the SRMR is defined as the ratio between the modulation energies in low and high modulation frequencies. Though reference-free, the SRMR actually uses an implicit reference, namely the modulation spectrogram of clean speech.

Lastly, [9] proposed an indicator based on the internal auditory representation, a kind of L^1-norm of the bi-spectra of the spectrograms of the envelop and the temporal fine structure, respectively, derived from the neurogram. This measure is well correlated with subjective scores. However, the use of an internal auditory model makes it computationally costly and unsuitable as a criterion for a speech enhancement algorithm.

Transposed to the field of numerical images, audio intelligibility could be compared to image sharpness. In image processing, several reference-free objective measures of sharpness are based on the importance of Fourier (or wavelet) phase in the perception of blur [10]. In particular, the global phase coherence (GPC, [11]) measures how the regularity of an image —defined by its total variation (TV)— is affected by the destruction of the phase information. The study presented in [12] proposes an new Sharpness Index (SI), computationally simpler but shown to behave similarly to the GPC. It measures the sensitivity of the TV to the convolution of the image by a white noise. It was successfully used as a criterion for blind image deblurring, outperforming methods based on TV minimization.

We here propose to transpose this principle to audio signals. A sharp image has a sparse gradient and this sparsity is reduced by phase randomization (GPC) or white noise convolution (SI), which increases the TV. On the contrary, the TV of a blurred or noisy image is much less sensitive to those operations. A similar behavior is encountered for audio signals. A clear sound has a sparse spectrogram, composed of thin segments (horizontal for harmonics, vertical for impulses), unlike a reverberated or noisy sound, where the segments are

smeared in time or frequency. Convolving the sound with a white noise should reduce the spectrogram sparsity for a clear sound, while letting it almost unchanged for a reverberated or noisy sound. Hence, we propose to define an audio sharpness index (aSI) as the sensitivity of the spectrogram sparsity to a convolution of the signal with a white noise. As image deblurring based on TV sensitivity maximization outperforms methods based on TV minimization, it is expected that replacing the L^1-norm criterion commonly used in audio enhancement algorithms (e.g. source separation) by the proposed aSI criterion would lead to better performances.

The paper is structured as follows. Through the analogy with the definition of the image sharpness index, we define an audio sharpness index in Section II. The proposed experimental validations relate to both a speech intelligibility measure (Section III) and audio enhancement algorithms (Section IV)[1].

II. The audio sharpness index

A. Spectrogram

Considering the time-frequency analysis of a finite-length discrete-time signal s, with analysis windows of length N overlapping of 50%, we define the spectrogram of s as

$$S(f,t) = \sum_{n=0}^{N-1} s(t+n)h(n)C(f,n) \quad (1)$$
$$\text{for} \quad f \in \{0,1,\ldots,N_f-1\} \quad \text{and} \quad t \in \frac{N}{2}\mathbb{Z},$$

where the apodization function h, the base functions C, and the value of N_f (N or $N/2$) depend on the transform used. In this work, we will only use real transforms, like the Discrete Cosine Transform.

The sparsity of the spectrogram will be measured by

$$\|S\|_1 = \sum_{f,t} |S(f,t)|. \quad (2)$$

B. Definition of the audio sharpness index

Let $s' = s * w$, where $*$ denotes the discrete convolution product and w is a white Gaussian noise with zero mean and variance σ_w^2. Let S and S' the spectrograms of s and s', respectively, as defined by Eq. (1). Inspired by the image Sharpness Index [12], we define the *audio Sharpness Index* of s as

$$aSI(s) \triangleq -\log\left(\Phi\left(\frac{\mathrm{E}[\|S'\|_1] - \|S\|_1}{\sqrt{\mathrm{Var}[\|S'\|_1]}}\right)\right), \quad (3)$$

where $\mathrm{E}[X]$ and $\mathrm{Var}[X]$ respectively denote the expectation and the variance of a random variable X, and

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-x^2/2}\, dx \quad (4)$$

is the tail of the normalized Gaussian distribution. Let us comment the definition given in Eq. (3). Ideally, we would like to compute the probability that the convolution of s with a white noise does not increase the sparsity of its spectrogram, that is,

$$p = \mathrm{Prob}\big[\|S'\|_1] \le \|S\|_1\big]. \quad (5)$$

This probability p is expected to be very small for a clean (and informative) audio signal, and not so small for a noisy and/or reverberated signal. Assuming that $\|S'\|_1$ is nearly Gaussian (which is observed in practice), we can estimate the quantity $-\log p$ (more adapted than p to a computer scale since values like $p = 10^{-10000}$ could be easily observed) by

$$-\log\left(\mathrm{Prob}\left[X \le \|S\|_1 \;\middle|\; X \sim \mathcal{N}\big(\mathrm{E}[\|S'\|_1], \mathrm{Var}[\|S'\|_1]\big)\right]\right),$$

which is exactly the quantity defined in Eq. (3).

In other terms, the audio Sharpness Index we defined in Eq. (3) measures the sensitivity of the spectrogram sparsity of a signal to the degradation caused by the convolution with a Gaussian white noise.

C. Computation

Theorem 1. *The expectations of $\|S'\|_1$ and $\|S'\|_1^2$ are*

$$\mathrm{E}[\|S'\|_1] = \sqrt{\frac{2}{\pi}} N_t \sum_{f=0}^{N_f-1} \sigma_{S'}(f) \quad (6)$$

$$\mathrm{E}[\|S'\|_1^2] = \frac{2}{\pi} \sum_{\substack{0\le f,f'\le N_f-1 \\ 1-N_t\le\Delta\le N_t-1}} (N_t - |\Delta|)\sigma_{S'}(f)\sigma_{S'}(f')\tilde{\omega}\left(\frac{\Gamma_{S'}(f,f',\frac{N}{2}\Delta)}{\sigma_{S'}(f)\sigma_{S'}(f')}\right) \quad (7)$$

where

- *N_t and N_f are the numbers of columns and lines of S;*
- *$\Gamma_{S'}(f,f',\tau) \triangleq \sigma_w^2 \mathrm{DCT}[\mathcal{R}_{s,\tau}(n,n')]$;*
- *$\mathcal{R}_{s,\tau}(n,n') \triangleq R_s(\tau+n-n')h(n)h(n')$, where R_s stands for the auto-correlation of s (finite and deterministic);*
- *$\sigma_{S'}^2(f) \triangleq \Gamma_{S'}(f,f,0)$;*
- *$\forall\, x \in [-1,1], \quad \tilde{\omega}(x) \triangleq x \arcsin x + \sqrt{1-x^2}$.*

One can then derive $\mathrm{Var}[\|S'\|_1]$ and compute $aSI(s)$ according to (3).

Proof. Convolving the deterministic finite-length signal s with the white noise w produces s' stationary, Gaussian with zero mean. Hence, $S'(f,t)$ is stationary too, and

$$\mathrm{E}[S'(f,t)] = 0 \quad (8)$$

$$\begin{aligned}&\mathrm{Var}[S'(f,t)] \\ &= \sum_{m,n=0}^{N-1} \mathrm{E}[s'(t+m)s'(t+n)]h(m)h(n)C(f,m)C(f,n) \\ &= \sigma_w^2 \sum_{m,n=0}^{N-1} R_s(m-n)h(m)h(n)C(f,m)C(f,n) \\ &\triangleq \tilde{\sigma}_{S'}^2(f) \quad (\text{independent of } t)\end{aligned} \quad (9)$$

Since $S'(f,t)$ is Gaussian and using Lemma 8 of [12],

$$\mathrm{E}[|S'(f,t)|] = \tilde{\sigma}_{S'}(f)\sqrt{\frac{2}{\pi}}, \quad (10)$$

[1]This work is part of the ICityForAll project, which was granted by the European program Ambient Assisted Living (AAL) from 2011 till 2015. See http://www.icityforall.eu

so that

$$\mathrm{E}[\|S'\|_1] = \sum_{f,t} \mathrm{E}[|S'(f,t)|] = \sqrt{\frac{2}{\pi}} N_t \sum_{f=0}^{N_f-1} \tilde{\sigma}_{S'}(f). \quad (11)$$

To obtain $\mathrm{E}[\|S'\|_1^2]$, we first compute

$$\begin{aligned}
&\mathrm{E}[S'(f,t)S'(f',t')] \\
&= \sum_{n,n'=0}^{N-1} \mathrm{E}[s'(t+n)s'(t'+n')]h(n)h(n')C(f,n)C(f',n') \\
&= \sigma_w^2 \sum_{n,n'=0}^{N-1} R_s(t-t'+n-n')h(n)h(n')C(f,n)C(f',n') \\
&= \sigma_w^2 \mathrm{DCT}[\mathcal{R}_{s,t-t'}(n,n')] \\
&= \Gamma_{S'}(f,f',t-t') \qquad (12)
\end{aligned}$$

Note that $\tilde{\sigma_{S'}}^2(f) = \Gamma_{S'}(f,f,0) = \sigma_{S'}^2(f)$, so that (11) is equivalent to (6). Moreover, using Lemma 9 of [12] with $Z = [S'(f,t), S'(f',t')]^\top$, we obtain

$$\mathrm{E}[|S'(f,t)S'(f',t')|] = \frac{2}{\pi}\sigma_{S'}(f)\sigma_{S'}(f')\tilde{\omega}\left(\frac{\Gamma_{S'}(f,f',t-t')}{\sigma_{S'}(f)\sigma_{S'}(f')}\right) \quad (13)$$

and Eq. (7) follows using

$$\mathrm{E}[\|S'\|_1^2] = \sum_{\substack{0\le f,f'\le N_f-1 \\ 0\le k,k'\le N_t-1}} \mathrm{E}\left[|S'(f,k\frac{N}{2})S'(f',k'\frac{N}{2})|\right]. \quad (14)$$

□

D. Parameter setting

Only two parameters must be set: the signal duration and the window duration in the spectrogram computation.

The latter is set around 20ms, as commonly used in audio processing. Taking longer windows makes the spectrogram less sensitive to smearing in the time dimension in case of reverberation, while shorter windows makes the spectrogram less sensitive to smearing in the frequency dimension in case of noise.

The signal duration T (and thus the width N_t of the spectrogram) must fulfill contradictory constraints. On one hand, T should be larger than the average syllable duration (*ca.* 400ms) to ensure the intra-speaker stability of the aSI. On the other hand, the computational cost increases with N_t, besides which T should be as small as possible in the foresight of using the aSI as criterion for non-stationary enhancement algorithms.

Finally, since silences and abrupt start or end in the signal are very sensitive to the convolution with the white noise, they would artificially increase the aSI. Consequently, silence suppression, fading in and fading out must be applied before computing the aSI.

III. Evaluation: the Audio SI as an intelligibility measure

A. Sound material

We used speech signals at 16 kHz from the TIMIT corpus [13]. We chose 16 speakers, one male and one female from each of the 8 dialect regions of the USA defined in the corpus documentation. For each speaker, the analyzed signal consists of the five SX sentences concatenated and has a duration of 9 to 18 s.

We computed the audio SI for each signal for various levels of noise and reverberation. For the reverberation, we considered a purely reverberant room impulse characterized by its reverberation time T60 (the time after which the sound has decreased of 60 dB below its original level). For each value of T60, we synthesized the impulse response by multiplying a white Gaussian noise by an exponential envelope matching T60. We added to the reverberated signal a white Gaussian noise at a given Signal to Noise Ratio (SNR). We tested 30 values of T60 logarithmically distributed between 10 ms and 5 s, and 21 values of SNR linearly distributed between -30 and +30 dB.

Hence, for each of the 16 speakers, we computed the aSI on the five-sentences signal, in each of the 630 (T60,SNR) conditions.

B. Audio SI computation

Before computation, we suppressed the silent parts in the signals. Then we computed the aSI on disjoint blocks of 512 ms, with the first and last 16ms attenuated by half Hamming windows. Finally, we computed the mean value of aSI on the whole signal. The spectrograms are based on 32 ms analysis windows.

C. Results

For each (SNR,T60) condition, we computed the average aSI over the 16 speakers. Fig 1 represents the iso-aSI lines in the SNR-T60 plane. This figure is very similar to that of the iso-STI lines (see [14]). To explore further this similarity, we plotted the relation aSI-STI for each triplet (speaker, SNR, T60) (see Fig. 2, where the STI is computed according to [14]). The log of the aSI is linearly correlated with the STI: the global correlation coefficient is 0.96, and the individual correlation coefficients of the speakers are between 0.94 and 0.98. This shows that the aSI can be considered as a good predictor of the STI, and thus used as an intelligibility measure, since the STI is generally considered as a reliable reference. However, the main advantage of the aSI is that it is non-intrusive, which represents a great practical interest.

IV. Evaluation: the Audio SI as an intelligibility criterion

A. Conceptual framework

We propose to experimentally validate the aSI as a criterion for blind source separation (BSS). Our idea is that a separated source is more intelligible than a mixture, so that, under the assumption that the aSI measures intelligibility, a separation algorithm could be driven by aSI maximization.

To demonstrate this idea, we will restrict to the simple case of a stereo instantaneous mixture of two sources. Let $\boldsymbol{s}$ a vector of two sources and A the mixture matrix, with $A(\cdot,i) = [\cos\theta_i \ \sin\theta_i]^T$. The mixture is $\boldsymbol{x} = A\boldsymbol{s}$. The goal of

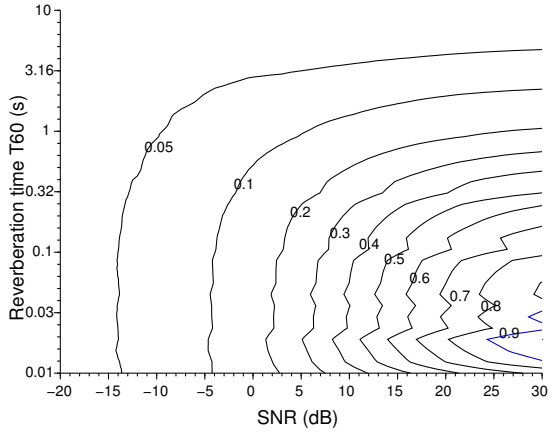

Fig. 1. Iso-aSI lines in the SNR-T60 plane, where for each (SNR,T60) condition, the aSI is averaged over the 16 speakers.

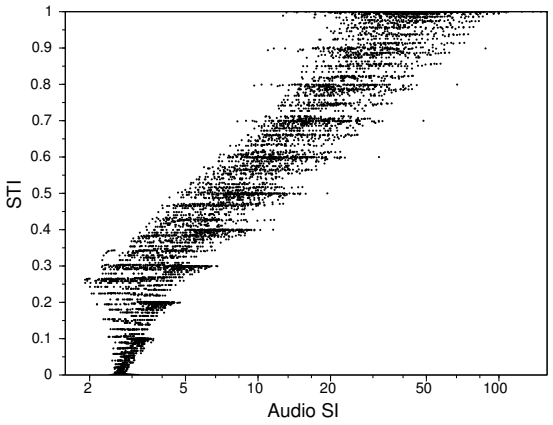

Fig. 2. Relation between aSI and STI: each point represents one condition (speaker,SNR,T60), where speaker = 1 to 16, T60 takes 30 logarithmically distributed values between 10 ms and 5 s, and SNR takes 21 linearly distributed values between -30 and +30 dB.

BSS is to estimate $\boldsymbol{s}$ from $\boldsymbol{x}$ with A unknown. If one estimates θ_1 and θ_2, the exact estimation is given by

$$y_i = \frac{1}{\sin(\theta_j - \theta_i)}(x_1 \sin\theta_j - x_2 \cos\theta_i), \quad i,j \in \{1,2\} \quad (15)$$

From Eq. (3) and Theorem 1, one can easily deduce that the aSI is invariant when the signal is multiplied by a scaling factor. Consequently, we estimate θ_1 and θ_2 by:

$$\{\hat{\theta}_1, \hat{\theta}_2\} = \arg\operatorname{local}\max_{\theta} \mathrm{aSI}(y_\theta) \,|\, y_\theta = x_1 \sin\theta - x_2 \cos\theta. \quad (16)$$

For the computation of $\mathrm{aSI}(y_\theta)$, replacing s by y_θ in the calculation of Section II leads to:

$$\begin{aligned}&\Gamma_{Y'_\theta}(f,f',\tau) = \\ &\sin^2(\theta)\Gamma_{X'_1}(f,f',\tau) + \cos^2(\theta)\Gamma_{X'_2}(f,f',\tau) \\ &-\sin(\theta)\cos(\theta)\big(\Gamma_{X'_1X'_2}(f,f',\tau) + \Gamma_{X'_2X'_1}(f,f',\tau)\big) \quad (17)\end{aligned}$$

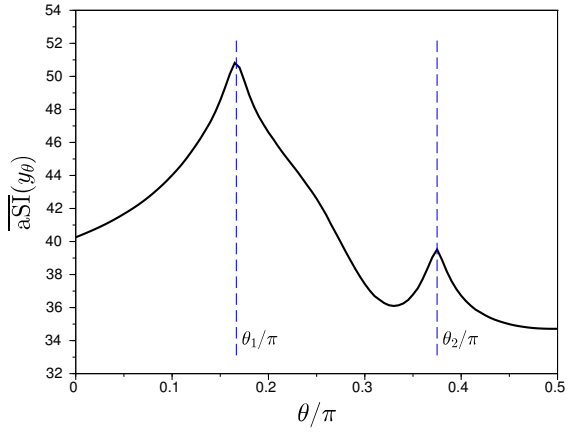

Fig. 3. Separation of a stereo mix of two speech signals: average aSI of the demixed signal y_θ as a function of the demixing parameter θ.

with the same notations as in Theorem 1 and:

- $\Gamma_{X'_iX'_j}(f,f',\tau) \triangleq \sigma_w^2 \mathrm{DCT}[\mathcal{R}_{x_i,x_j,\tau}(n,n')]$;
- $\mathcal{R}_{x_i,x_j,\tau}(n,n') \triangleq R_{x_i,x_j}(\tau+n-n')h(n)h(n')$, where R_{x_i,x_j} stands for the inter-correlation of x_i and x_j.

Hence, once $\Gamma_{X'_1}(f,f',\tau)$, $\Gamma_{X'_2}(f,f',\tau)$, $\Gamma_{X'_1X'_2}(f,f',\tau)$, and $\Gamma_{X'_2X'_1}(f,f',\tau)$ have been computed, $\mathrm{aSI}(y_\theta)$ can be easily computed for any value of θ, using Eq. (3) and Theorem 1.

B. Experiment 1: separation of two speech signals

We mixed two speech signals, one from a male speaker, the other from a female speaker, from the TIMIT corpus. The shorter signal was extended to the length of the longer one by zero-padding, leading to a signal duration of 2.9 s. We set $\theta_1 = \pi/6$ and $\theta_2 = 3\pi/8$.

For $\theta = 0$ to $\pi/2$, by step of $\pi/200$, and y_θ defined by Eq. (16), we computed $\mathrm{aSI}(y_\theta)$ by blocks of 512 ms (with spectrograms based on 32 ms analysis windows) and averaged the aSI on the whole signal, resulting in $\overline{\mathrm{aSI}}(y_\theta)$. Note that in this case, silence-suppression and fading are not necessary, since the goal is to maximize $\overline{\mathrm{aSI}}(y_\theta)$ on θ, and not to measure the intelligibility.

Figure 3 shows $\overline{\mathrm{aSI}}(y_\theta)$ as a function of θ. The maxima match exactly θ_1 and θ_2, which allows a perfect separation. Note that $\overline{\mathrm{aSI}}(y_{\theta_1}) = \overline{\mathrm{aSI}}(s_2)$ and $\overline{\mathrm{aSI}}(y_{\theta_2}) = \overline{\mathrm{aSI}}(s_1)$.

C. Experiment 2: separation of two music signals

In this paper, we have focused on speech, but the definition of aSI does not restrict to speech signals: it is suitable for general audio signals. We repeated the previous experiment with a 13s singing voice and its piano accompaniment, extracted from the QUASI database [15], [16] and re-sampled at 32 kHz. We used 256ms-blocks for aSI computation.

As illustrated by Fig. 4, the average aSI is maximum for θ_2, which extracts the voice (s_1), but it is minimum for θ_1, which extracts the piano (s_2). This can be explained by the large difference between voice aSI (14) and piano aSI (1.3).

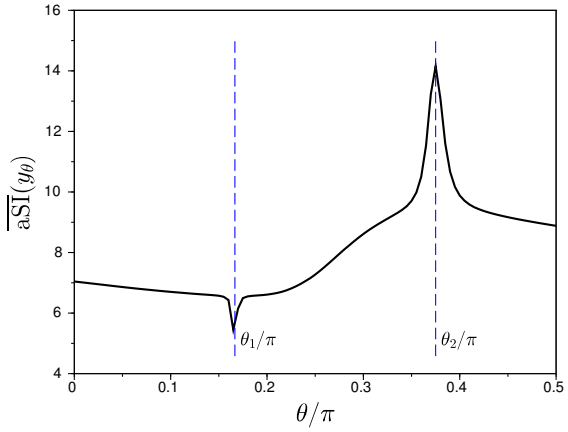

Fig. 4. Separation of a stereo mix of a singing voice and its piano accompaniment: average aSI of the demixed signal y_θ as a function of the demixing parameter θ.

TABLE I
PERFORMANCE OF SOURCE SEPARATION BY aSI MAXIMIZATION, COMPARED TO THAT OF FastICA. SDR = SOURCE TO DISTORTION RATIO, SIR = SOURCE TO INTERFERENCE RATIO, SAR = SOURCE TO ARTIFACT RATIO, MEASURED ACCORDING TO [19]

		SDR	SIR	SAR
voice	FastICA	37	37	73
	Max aSI	70	109	70
piano	FastICA	34	34	67
	Max aSI	43	43	67

Consequently, the assumption that a separated source has a larger aSI than a mixture does not hold anymore here.

In this case, the sources can be successively extracted using iterative deflation [17]: we first extracted the voice signal by maximizing $\overline{\mathrm{aSI}}(y_\theta)$, then we estimated its contribution to the mixture, finally we estimated the piano signal by subtracting this contribution.

The estimation error on θ_1 and θ_2 is less than $\pi/1000$. As indicated by Table I, our separation method outperforms the classical method FastICA [18] on this simple example. An audio demonstration is available at
http://www.mi.parisdescartes.fr/%7Emahe/Recherche/audioSI/

V. Conclusion

We have advanced and expressed as a closed-form formula a new non-intrusive speech intelligibility measure, the audio Sharpness Index (aSI). It is defined as the sensitivity of the spectrogram sparsity to the convolution of the signal with a white noise. Our experiments have validated the aSI both as an intelligibility measure and as an intelligibility/clarity criterion to drive blind source separation (BSS).

As an intelligibility measure, the advantage of the aSI is not only its non-intrusiveness, but also the fact that it does not rely on any implicit clean reference, unlike [8].

Further experiments of sound enhancement have to be considered to complete the validation of the aSI as an intelligibility/clarity criterion. However, BSS based on aSI maximization has already a great advantage: it does not need any classical assumption of independence and non-Gaussianity of the sources.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Bosa Roca, United States: Taylor & Francis Inc., 2007, ch. 10. Evaluating performance of speech enhancement algorithms.

[2] *Methods for calculation of the speech intelligibility index*, ANSI Std. S3.5-1997, 1997.

[3] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[5] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb 2014.

[6] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. v. Waterschoot, and P. A. Naylor, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 719–732, April 2016.

[7] D. Sharma, P. A. Naylor, and M. Brookes, "Non-intrusive speech intelligibility assessment," in *21st European Signal Processing Conference (EUSIPCO 2013)*, Sept 2013.

[8] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sept 2010.

[9] M. E. Hossain, W. A. Jassim, and M. S. A. Zilany, "Reference-free assessment of speech intelligibility using bispectrum of an auditory neurogram," *PLoS ONE*, vol. 11, no. 3, march 2016.

[10] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[11] G. Blanchet, L. Moisan, and B. Rouge, "Measuring the global phase coherence of an image," in *2008 15th IEEE International Conference on Image Processing*, Oct 2008, pp. 1176–1179.

[12] A. Leclaire and L. Moisan, "No-reference image quality assessment and blind deblurring with sharpness metrics exploiting Fourier phase information," *Journal of Mathematical Imaging and Vision*, vol. 52, no. 1, pp. 145–172, 2015.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993.

[14] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[15] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[16] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3155–3167, 2011.

[17] P. Comon and P. Jutten, *Handbook of Blind Source Separation*, P. Comon and P. Jutten, Eds. Academic Press, 2010.

[18] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[19] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.