DESIGNING MULTICHANNEL SOURCE SEPARATION BASED ON SINGLE-CHANNEL SOURCE SEPARATION

A. Ramírez López[⋆]*, N. Ono*[†,‡]*, U. Remes*[⋆]*, K. Palomäki*[⋆]*, M. Kurimo*[⋆]

[⋆]Department of Signal Processing and Acoustics, Aalto University School of Electrical Eng., Finland
[†]Principles of Informatics Research Division, National Institute of Informatics, Japan
[‡] The Graduate University for Advanced Studies (SOKENDAI), Japan

ABSTRACT

In this paper, an extension of independent vector analysis (IVA), model-based IVA, is proposed for multichannel source separation. For obtaining better source models, we introduce a single-channel source separation method, and utilize the outputs as source variances in time-frequency-variant Gaussian source model. The demixing matrices are estimated in the same way as a state-of-the-art IVA method, auxiliary-function-based IVA (AuxIVA). Experimental evaluations show that the proposed approach is effective and improves the source separation performance of IVA. In addition, several post-filters aiming to realize multichannel Wiener filter (MWF) are investigated. This setup proves to further increase the performance of IVA. The presented method shows a potential to provide a general way to improve the separation performance from single-channel source separation to multichannel source separation.

Index Terms— independent vector analysis, blind source separation, speech source model, speech enhancement

1. INTRODUCTION

Blind source separation (BSS) is a technique to extract desired sources from mixtures with no knowledge of either the mixing process or the sources. In the convolutive overdetermined case, independent component analysis (ICA) in the frequency domain [1] has been a common source separation method. In ICA, source separation is obtained by seeking for statistically independent sources, which are represented via statistical source models. More recently, a multivariate variant of ICA has been developed, which is independent vector analysis (IVA) [2, 3, 4], where all the frequency components are modelled as stochastic vector variables and the sources are separated vector-wise instead of frequency-wise, as it occurs in ICA. IVA is an advantageous approach since, theoretically, it avoids permutation ambiguity due to the dependencies over the spectral channels represented in the source model.

In conventional IVA, an identical source model, typically a spherical multivariate super-Gaussian distribution, is assumed for all sources [2, 3]. However, it is not correct in several cases. For example, in speech and noise separation tasks, the sources' spectra have quite different characteristics. Speech spectrum is temporally non-stationary and has a structure caused by pitch and formants, while surrounding noise has broad band spectrum and may be temporally stationary. Therefore, the common source model in IVA does not reflect the differences of source characteristics between speech and noise, and consequently the separation performance could be insufficient. Another problem is diffuse noise. It is well known that using only a multichannel linear filter (beamformer) is not sufficient to suppress diffuse noise, so post-filtering can improve its performance [5]. BSS methods also have a limited capability to reduce diffuse noise.

Based on these motivations, we introduce in this work single-channel source separation as a module to provide a better source model in IVA. Including spectral subtraction, most single-channel source separation methods utilize the difference of spectrogram characteristics in each source. Hence, in this work, the source models are time-frequency-variant Gaussian distributions, which is similar to non-negative matrix factorization with Itakura-Saito divergence [6] or its multichannel version [7, 8]. From now on, this IVA extension will be called model-based IVA. In this work, the source model variances are computed from spectral subtraction, but we must note, to emphasize generality of our approach, that the variances could be provided from any other single-channel source separation method. Moreover, we also discuss how to design a post-filter based on the same single-channel source separation approach. The proposed method, model-based IVA, is evaluated without and with post-filtering, in a two-channel speech and noise separation task.

2. FREQUENCY-DOMAIN BSS

BSS in frequency-domain can be formulated as follows. We assume here that M source signals are observed by M microphones, and their short-time Fourier transform (STFT) representations are obtained. In the frequency-domain approach for convolutive mixtures, dependencies between the source signals and observed mixtures is modeled as a linear mixing process:

$$\boldsymbol{X}_{\tau\omega} = A_\omega \boldsymbol{S}_{\tau\omega}, \tag{1}$$

where $\boldsymbol{X}_{\tau\omega} = [X_{1\tau\omega}, \ldots, X_{M\tau\omega}]^T$ denotes the $M \times 1$ observation vector and $\boldsymbol{S}_{\tau\omega} = [S_{1\tau\omega}, \ldots, S_{M\tau\omega}]^T$ the $M \times 1$ source vector at frequency channel ω in time frame τ, and A_ω is the unknown mixing matrix associated with channel ω. The vector component $X_{m\tau\omega}$ denotes the mixture observed with microphone m and $S_{m\tau\omega}$ the mth source signal at channel ω at time frame τ. The estimated source signals $\boldsymbol{Y}_{\tau\omega}$ are computed by the linear demixing process:

$$\boldsymbol{Y}_{\tau\omega} = W_\omega \boldsymbol{X}_{\tau\omega}, \tag{2}$$

where $\boldsymbol{Y}_{\tau\omega} = [Y_{1\tau\omega}, \ldots, Y_{M\tau\omega}]^T$ and W_ω is the demixing matrix. The source separation problem involves finding W_ω based on observations $\boldsymbol{X}_{\tau\omega}$.

3. EXTENSION OF INDEPENDENT VECTOR ANALYSIS

3.1. Source Model of Conventional Independent Vector Analysis

In IVA, the sources to be separated are modelled by means of statistical source models. Besides, the dependency between the spectral channels of each source is represented in IVA by using a multivariate probability density function $p_y(\tilde{\boldsymbol{Y}}_{m\tau})$ as source model, for a source-wise vector $\tilde{\boldsymbol{Y}}_{m\tau} = [Y_{m\tau 1}, \ldots, Y_{m\tau\Omega}]^T$, where Ω is the total number of spectral channels and m is the number of source channel. Conventionally, IVA methods use spherical, time-invariant, and super Gaussian distributions [2, 3], such as

$$p_y(\tilde{\boldsymbol{Y}}_{m\tau}) \propto \exp\left\{-K\sqrt{\|\tilde{\boldsymbol{Y}}_{m\tau}\|_2^2}\right\}, \tag{3}$$

where K is a time-invariant constant and $\|\cdot\|_2$ denotes the L_2 norm of a vector.

3.2. Model-based Independent Vector Analysis

IVA has been previously evaluated with time-variant source distributions in [9, 10], where distribution variances were assumed constant across frequency channels. In the present paper, we propose a time-frequency-variant Gaussian distribution such as

$$p_y(Y_{m\tau\omega}) \propto \frac{1}{\sigma^2_{m\tau\omega}} \exp\left\{-\frac{Y^2_{m\tau\omega}}{\sigma^2_{m\tau\omega}}\right\}, \tag{4}$$

where $\sigma^2_{m\tau\omega}$ is the variance of mth source at time frame τ and frequency ω. In this work, we assume that we have a single-channel source separation method which separates a single channel observation into M source estimates. Then, the variances $\sigma^2_{m\tau\omega}$ are calculated as,

$$\sigma^2_{m\tau\omega} = |\hat{S}_{m\tau\omega}|^2, \tag{5}$$

where $\hat{S}_{m\tau\omega}$ is the output from single-channel source separation for the mth source at time frame τ and frequency ω.

3.3. Objective Function of Independent Vector Analysis

In IVA, the demixing matrices are iteratively estimated by minimizing the following objective function over W_ω,

$$J_1 = \sum_m \frac{1}{T}\sum_\tau G(\tilde{\boldsymbol{Y}}_{m\tau}) - \sum_\omega \log\det|W_\omega| \tag{6}$$

This function is derived from the Kullback-Leibler divergence between the p.d.f of the observed signal and that of the source model [2, 3, 4]. $G(\tilde{\boldsymbol{Y}}_{m\tau})$ is called *contrast function* and it is computed as $G(\tilde{\boldsymbol{Y}}_{m\tau}) = -\log p_y(\tilde{\boldsymbol{Y}}_{m\tau})$, where $p_y(\tilde{\boldsymbol{Y}}_{m\tau})$ is the multivariate p.d.f of the source model. T is the total number of frames. The minimization of (6) is equivalent to maximum likelihood (ML) estimation.

Given the proposed source model in (4), the objective function J_1 can be rewritten as,

$$J_2 = \sum_\omega \left(\sum_m \frac{1}{T}\sum_\tau \frac{\|\boldsymbol{w}^H_{m\omega}\boldsymbol{X}_{\tau\omega}\|_2^2}{\sigma^2_{m\tau\omega}} - \log\det|W_\omega|\right), \tag{7}$$

which is the objective function of model-based IVA, where $\boldsymbol{w}^H_{m\omega}$ is the mth row of the demixing matrix W_ω and ^{H} denotes Hermitian transpose.

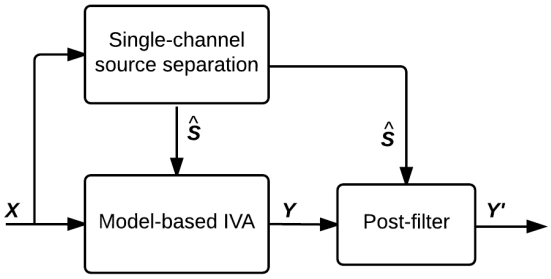

Fig. 1. Block diagram of the single and multichannel source separation system with post-filter setup proposed. $\boldsymbol{X}$ denote the observation vector and $\boldsymbol{Y'}$ the estimated sources vector.

3.4. Update Rules for Demixing Matrix

Traditionally, IVA algorithms have as standard solution the natural gradient update [2, 3, 4]. However, this kind of solution involves a trade-off between convergence speed and stability. New, more effective update rules based on auxiliary function technique were developed first for ICA [11] and later extended to IVA with the AuxIVA method [12].

The AuxIVA method involves two alternative update steps. In the extension of AuxIVA with the new source model proposed, the update rules are as follows. First, the weighted covariance matrices $V_{m\omega}$ are once calculated for all ω as

$$V_{m\omega} = \frac{1}{T}\sum_\tau \left(\frac{\boldsymbol{X}_{\tau\omega}\boldsymbol{X}^H_{\tau\omega}}{\sigma^2_{m\tau\omega}}\right) \tag{8}$$

Then, the demixing matrices are updated. Note that a closed-form solution for updating $\boldsymbol{w}_{m\omega}$ in eq. (7) simultaneously has not been proposed yet. Instead, we consider an update of only $\boldsymbol{w}_{m\omega}$ while keeping other $\boldsymbol{w}_{l\omega}(l \neq m)$ fixed. Therefore, the update rules for demixing matrix, applied for all ω, are:

$$\boldsymbol{w}_{m\omega} \leftarrow (W_\omega V_{m\omega})^{-1}\boldsymbol{e}_m, \tag{9}$$

$$\boldsymbol{w}_{m\omega} \leftarrow \frac{\boldsymbol{w}_{m\omega}}{\sqrt{\boldsymbol{w}^H_{m\omega}V_{m\omega}\boldsymbol{w}_{m\omega}}}, \tag{10}$$

where $\boldsymbol{e}_m$ is a unit vector with the mth element unity $\boldsymbol{e}_m = [0, \ldots, 1, \ldots, 0]$. The update rules are applied iteratively until convergence is achieved.

4. POST-FILTER DESIGN

Experiments conducted on MVDR beamformers indicate that a single-channel Wiener post-filter can improve their source separation performance [5]. In the current work, we present an analogous setup, where the single-channel-based multichannel source separation system proposed, model-based IVA, is concatenated with a time-variant post-filter. This setup is presented in Figure 1. The sources estimates $Y'_{m\tau\omega}$ are calculated based on the multichannel estimates $Y_{m\tau\omega}$ from model-based IVA as $Y'_{m\tau\omega} = H_{m\tau\omega}Y_{m\tau\omega}$, where $H_{m\tau\omega}$ is the STFT representation of the post-filter applied on the mth source estimate in time frame τ and frequency channel ω. In this work, we evaluate three time-variant post-filters $H_{m\tau\omega}$ calculated based on the multichannel source estimates $Y_{m\tau\omega}$ and source estimates $\hat{S}_{m\tau\omega}$ calculated with a single-channel source separation.

Wiener filter approach 1

The optimal post-filter in the minimum mean square error (MMSE) sense is obtained as Wiener filter, which is generally represented as ϕ_{ss}/ϕ_{yy} where ϕ_{ss} and ϕ_{yy} are the expectation of the power of the target signal s and the observation y, respectively. In the first post-filter, we calculate it as

$$H^{(1)}_{m\tau\omega} = \frac{|\hat{S}_{m\tau\omega}|^2}{|\hat{S}_{m\tau\omega}|^2 + |N_{m\tau\omega}|^2}, \tag{11}$$

where the noise $N_{m\tau\omega}$ is calculated as $N_{m\tau\omega} = Y_{m\tau\omega} - \hat{S}_{m\tau\omega}$.

Wiener filter approach 2

In the second post-filter, the expectation of the observation power in Wiener filter is directly calculated as the power of the observation. For guaranteeing that the post-filter value falls within the range from 0 to 1, simply clipping is applied. Then, it can be represented as

$$H^{(2)}_{m\tau\omega} = \min\left\{\frac{|\hat{S}_{m\tau\omega}|^2}{|Y_{m\tau\omega}|^2}, 1\right\}. \tag{12}$$

Amplitude replacing

Most single-channel source separation techniques estimate the sources via a time-frequency mask with the original observed phase. In contrast, model-based IVA is a multichannel linear filter that estimates not only the amplitude but also the phase. In the third approach, we try to combine the amplitude estimation from single-channel source separation and the phase estimation from model-based IVA. This is achieved with the following post-filter:

$$H^{(3)}_{m\tau\omega} = \frac{|\hat{S}_{m\tau\omega}|}{|Y_{m\tau\omega}|}. \tag{13}$$

Applying single source separation to each of model-based IVA outputs

As the last noise suppression approach evaluated in this work, we apply the single-channel source separation technique to each output of model-based IVA as a post-filter.

5. EXPERIMENTAL EVALUATIONS

5.1. Data

We evaluated our methods on part of the material from the Signal Separation Evaluation Campaign (SISEC) 2013 [13]. We used in particular the development set of the two-channel mixtures of speech and real-world background noise task. This set consists of nine stereo recordings of a speech source that is contaminated by real-world diffuse noise. The diffuse noise was recorded in three kinds of public environments: a subway car, cafeterias and squares. Apart from the nine stereo mixtures, the set also includes the corresponding source signals (speech and noise) and *source images*, which are the convolved versions of each separate source signal observed at the microphones. All signals of this data set have a duration of 10 s. and their sampling frequency is 16000 Hz.

5.2. Setup

The system proposed in this work uses a single-channel source separation method to improve multichannel source separation performance. The single-channel source estimates used in this work were calculated with the spectral subtraction method implemented in VOICEBOX [14]. The method was used with the preset parameters. The source estimates calculated with spectral subtraction were used to provide source variances to model-based IVA that was implemented on AuxIVA [12]. AuxIVA was applied on STFTs calculated in 2048-sample Hamming windows with 50% overlap. An identity matrix was used as initial value for the demixing matrix, and the algorithm was iterated 20 times to ensure convergence.

The proposed single-channel-based multichannel source separation system is evaluated without a post-filter and with the three post-filters proposed in Section 4. For comparison, we also evaluate the proposed system performance when the post-filter is substituted with spectral subtraction. Since the system is evaluated in a speech enhancement task, post-filtering is applied to the output of model-based IVA corresponding to the speech source. The output corresponding to the diffuse noise source is discarded.

5.3. Evaluation

The speech enhancement task was evaluated using the Signal to Distortion Ratio (SDR_i) computed with the BSS Eval Matlab toolbox [15, 16]. This measure evaluates estimated source images and it was one of the evaluation metrics of SISEC 2013. However, this kind of energy ratio's evaluation criteria cannot explain certain auditory properties [17]. Therefore, the experiments were also evaluated with other measure, frequency-weighted segmental signal-to-noise ratio (fwSNRseg) [18]. This measure has proven to have high correlation with subjective assessments on speech quality [19].

Estimates of the stereo speech source images were evaluated by averaging over the two channels. The final SDR_i and fwSNRseg values were obtained after averaging over all trials within each of the three noise background conditions and finally averaging over these three cases.

5.4. Results

The systems evaluated in the current work include baseline single-channel (spectral subtraction) and multichannel (conventional IVA) source separation systems, and the proposed single-channel-based multichannel source separation system, model-based IVA, which is evaluated by itself, with several post-filters and with spectral subtraction. The results are presented in Figure 2. We can see that the baseline single-channel system performance with spectral subtraction is better than the baseline multichannel system performance with conventional IVA. The multichannel system performance improves when the single-channel estimates are used as source variances in model-based IVA. Evaluation with the fwSNRseg measure suggests that model-based IVA performance was better than spectral subtraction, SDR_i favoured spectral subtraction. This is because the evaluation measures emphasise different qualities in the separated signal. Listening to the audio samples indicated that spectral subtraction removes more background noise than model-based IVA but introduces audible distortion in the speech signal.

Model-based IVA performance improves when the output signal is post-filtered, and both evaluation measures indicate that model-based IVA performance with post-filter $H^{(1)}$ is better than model-based IVA or spectral subtraction performance. Model-based IVA performance with post-filter $H^{(1)}$ is close to the performance of

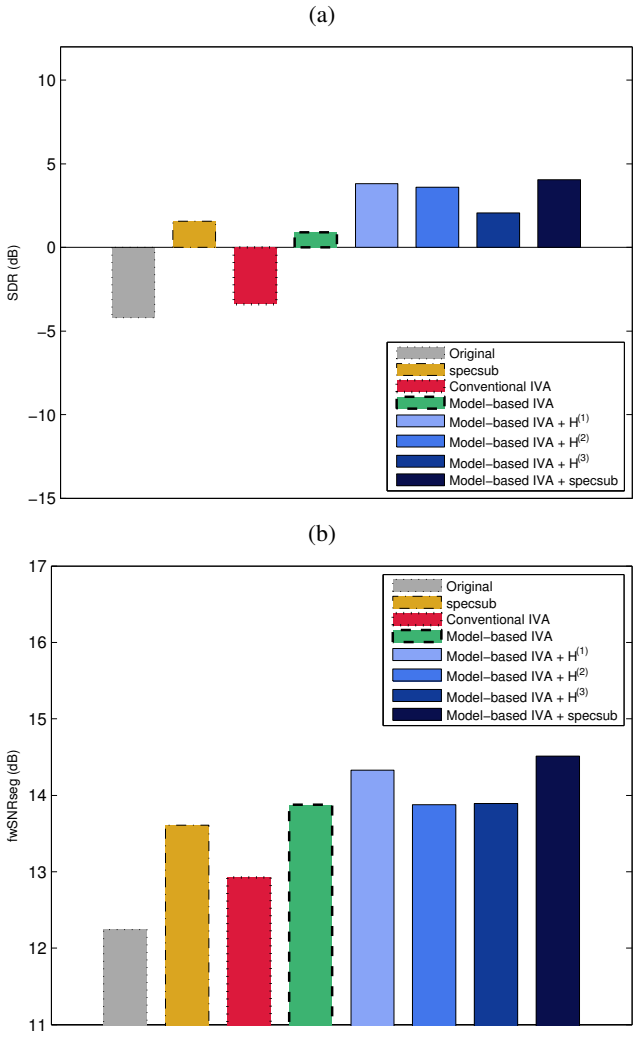

Fig. 2. SDR_i and fwSNRseg results of the original mixtures and spectral subtraction technique (for reference), and the methods under comparison: conventional IVA, model-based IVA and model-based IVA with post-filters $H^{(1)}$, $H^{(2)}$ and $H^{(3)}$; and model-based IVA with spectral subtraction post-processing.

model-based IVA with spectral subtraction post-processing. Based on the differences between SDR_i and fwSNRseg results, it seems that both Wiener filters $H^{(1)}$ and $H^{(2)}$ are more efficient in noise reduction than $H^{(3)}$.

6. DISCUSSION

In this paper, we presented model-based IVA that extends conventional IVA methods with a time-frequency-variant Gaussian source model. The source variances were calculated based on single-channel source separation outputs. The model-based IVA was evaluated in a speech enhancement task with two-channel speech and noise mixtures. The comparison between conventional IVA and model-based IVA validated the hypothesis that IVA performance in source separation can be improved by using improved source models even though they are provided by a simple method such as spectral subtraction. While the current work focused on using single-channel source separation to improve multichannel source separation, the source variances need not be determined in this manner. Future work on source model variances or better source models also can further improve IVA performance.

The multichannel system proposed in this work was completed with a single-channel post-filter. Model-based IVA was evaluated with three post-filters; all of them improved the performance of model-based IVA. The post-filtering improvement is more prominent with SDR_i. This suggests that the post-filtering solutions proposed improve the source separation performance of model-based IVA by increasing the amount of background noise removed. The best performance over the three post-filters was observed with $H^{(1)}$. The best overall performance was obtained when spectral subtraction was applied on the model-based IVA output. However, to achieve this result, we had to apply spectral subtraction twice in this case: first to the two input signals and then to the output calculated with model-based IVA. Since spectral subtraction does not need to be applied on the output when post-filters are used, the post-filter approach is more efficient computationally.

7. ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland by the grants no 136209, 272710 and Center of Excellence in Computational Inference (grant no 251170), and by a Grant-in-Aid for Scientific Research (A) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 23240023).

8. REFERENCES

[1] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[2] T. Kim, T. Eltoft, and T-W. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components," in *Proc. ICA*, 2006, pp. 165–172.

[3] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA, Using Multivariate Probability Density Functions," in *Proc. ICA*, 2006, pp. 601–608.

[4] T. Kim, H. T. Attias, S-Y. Lee, and T-W. Lee, "Blind Source Separation Exploiting Higher-order Frequency Dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[5] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering Techniques," in *Microphone Arrays*, pp. 39–60. Springer Berlin Heidelberg, 2001.

[6] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[7] A. Ozerov and C. Févotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[8] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel Extensions of Non-negative Matrix Factorization with Complex-valued Data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[9] T. Ono, N. Ono, and S. Sagayama, "User-guided Independent Vector Analysis with Source Activity Tuning," in *Proc. ICASSP*, 2012, pp. 2417–2420.

[10] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA*, 2012.

[11] N. Ono and S. Miyabe, "Auxiliary-function-based Independent Component Analysis for Super-Gaussian Sources," in *Proc. LVA/ICA*, 2010, pp. 165–172.

[12] N. Ono, "Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[13] N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito, "The 2013 Signal Separation Evaluation Campaign," in *Proc. MLSP*, 2013, pp. 1–6.

[14] M. Brookes, "VOICEBOX: A Speech Processing Toolbox for MATLAB," 2006.

[15] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Proc. ICA*, 2007, pp. 552–559.

[16] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A Study of Complexity and Quality of Speech Waveform Coders," in *Proc. ICASSP*, 1978, pp. 586–590.

[19] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. ASLP*, vol. 16, no. 1, pp. 229–238, 2008.