A NEW TIME-FREQUENCY APPROACH FOR UNDERDETERMINED CONVOLUTIVE BLIND SPEECH SEPARATION

Mariem Bouafif [1], *Zied Lachiri* [1]

[1] Signal, Image and Information technology Laboratory, National Engineering School of Tunis (ENIT)

ABSTRACT

In this paper, we present a new Time-Frequency approach for recovering sources' contribution to two convolutive mixtures. The separation task is performed on two steps: Each mixture is clustered into Voiced/Unvoiced frames, and then the predominant source in each time frequency bin is identified through a specific weight function which is based on sources' excitation characteristics extraction. We investigate the performance of the proposed approach in the underdetermined context using objective quality measures. Results for separating three and four speech sources in a live recorded mixture show the superiority of the proposed method in rejecting artifacts over existing convolutive separation techniques.

Index Terms— Blind Source Separation, Glottal Closure Instants, Time Delay of Arrival, Weight Function

1. INTRODUCTION

Recovering original sources from a set of mixed observations is the ability to relatively emphasize desired source from a mixture in a multi-speakers environment, like meetings, conferences, and cocktail parties. This task is named Blind Source Separation (BSS) [1]-[3] when assuming *no apriori* informations. It attempts to extract the original unknown speech source signals from their mixtures using only observed information without knowledge about the number of sources or the mixing process.

BSS can be achieved appropriately using Independent Component Analysis (BSS/ICA) [4]-[10] which is straightforward only for determined and overdetermined instantaneous mixtures. However in the convolutive underdetermined case, Time-frequency masking [11]-[20], or Maximum a Posteriori (MAP) estimation [21] are widely employed. These approaches perform separation by exploiting sparsity in the Time-Frequency domain. However, their performance is still limited by temporal fluctuations which considerably introduce artifacts in the estimated speech. Some other approaches which rely on specific features' extraction [11]-[18], are known by their simplicity and effectiveness in the determined context and only under low reverberation.

In general, BSS algorithms are always focusing on distortion and interferences rejection. However, optimal performance cannot be addressed without dealing with artifact suppression.

Motivated by this shortcoming, this work presents a new Time-Frequency separation method dealing with convolutive speech mixtures. It is a two stages approach: firstly, observed mixture are clustered into Voiced/ Unvoiced regions, and secondly voiced regions are classified as originating from the same source referring to a specific weight function values assuming sources' sparseness. This task is based on Glottal Closure Instants detection referring to sources' Time Delay of Arrival (TDOA). This method is inspired from an existing Temporal and Spectral Processing (TSP) technique [22]. In fact, our new technique is employed on the input observed mixture instead of the temporally processed speech signal already used in TSP. As a result, we improve rejection of artifact or musical noise compared to previous method. Our approach is in line with TSP which is developed only for the determined context. However, our technique is extended to fit the underdetermined context. In that sense, the proposed approach is more general. Moreover, the benefit of this technique is that it is robust against musical noise, by proving its effectiveness in rejecting artifact compared to the latest BSS technique.

The reminder of this paper is organized as follows: Section 2 provides an overview of the convolutive mixing model and states the problem. Section 3 describes the proposed method. Section 4 explains the experimental evaluation and presents a discussion on the achieved results. Finally, section 5 concludes the paper.

2. CONVOLUTIVE MIXING MODEL: OVERVIEW AND PROBLEM STATEMENT

We consider a two- microphones array in reverberant environment where I speech sources are present. A convolutive mixing model can be assumed, by which the observation at the j^{th}microphone x_j can be modeled as a summation of individual contribution by I involved sources. Hence, the convolutive mixing model is expressed as follows:

$$x_j(n) = \sum_{i=1}^{I} s_{ij}^{img}(n) + v_j(n) \quad (1)$$

Where $s_{ij}^{img}(n)$ is the spatial image of the i^{th} source on the j^{th} channel. It is the contribution of the i^{th} source to the j^{th} mixture. v_j denotes additive noise.

In a convolutive mixture, the spatial image of source is generally expressed as follows:

$$s_{ij}^{img}(n) = \sum_n h_{ij}(n)\ s_i(n - \tau_i) \quad (2)$$

Where h_{ij}, denotes mixing coefficient filter modelling the acoustic path from the i^{th}source signal to the j^{th} microphone and τ_i is the Time Delay of Arrival (TDOA) of the i^{th}source signal.

The main goal of any BSS system is to recover either the original signals s_i or their spatial images s_{ij}^{img} given J mixture channels. In the proposed approach, we focus on extracting a set of estimated sources' images A as follows:

$$A = \begin{Bmatrix} \widehat{s_{11}^{img}} & \cdots & \widehat{s_{1J}^{img}} \\ \vdots & \cdots & \vdots \\ \widehat{s_{I1}^{img}} & \cdots & \widehat{s_{IJ}^{img}} \end{Bmatrix} \quad (3)$$

Where each set corresponds to each of the involved source in the mixture $\widehat{s_{ij}^{img}}$ is an estimation of the i^{th} spatial image source signal s_{ij}^{img} at the j^{th} microphone. A is determined using only two microphones ($J = 2$), and without any prior guess on sources s_i.

In the following, we detail the proposed case approach of three simultaneously speaking speech sources.

3. THE PROPOSED APPROACH

Our approach is based on a two-stage structure: In the first stage, it performs Voice/Unvoice detection and it synthesizes speech signal in the second stage with a weight filter. The rest of this section details the processing steps.

3.1. Voiced/Unvoiced Clustring

The microphone observations are framesized into Voiced/ Unvoiced frames, using a Voiced/Unvoiced Decision (VUD) algorithm based on pitch detection.

The observed mixed signal is firstly frame-sized into blocks of 40ms overlapped by 10ms, and then subjected to a normalized autocorrelation [23]. Then we take the half of the autocorrelation of each block, as it's just a mirror for a real signal. As the human pitch F_0 is in the range [50 Hz, 500Hz], we seek the correlation sequence over the lag range [2ms: 20ms].

Each speech frame, subjected to autocorrelation, is clustered into voiced or unvoiced frames depending on two parameters: the first major peak (R_p) [23], and the similarity behaviour (S) of the autocorrelation frame [24]. In a perfect anechoic condition, each frame is voiced when the first major peak $R_p = 1$, and its similarity behaviour $S = 1$. However, in a practical condition each frame is considered as voiced only if the first major peak $R_p \geq 0.4$, and it's similarity behaviour $S \geq 0.7$. We should note that these thresholds are chosen referring to a computed weight filter's amplitude in each voiced frame under reverberant conditions where $RT_{60} \leqslant 250ms$

Knowing that the extracted pitch in a voiced region can be produced by any involved source in a multi-speakers mixture, we need another feature to specify the active speaker in each voiced frame. This stage is performed by estimating a weight function which enhances predominant speaker from other involved sources.

In the next subsection, we detail the design of a weight function and its use to synthesize estimated spatial source's image.

3.2. Spatial sources images' construction

The estimation of the spatial image of each involved source in the observed mixture is performed without any prior guess about the number of involved sources. However, sources counting is essential in the speech sources separation task.

In this paper our sources' counting algorithm depends on sources' Time Delay of Arrival (TDOA's) estimation. We propose the use of our sparseness based TDOA estimator technique, which exploits the pseudo-periodicity of sources' instants of significant excitation (GCI's). We have previously shown that this technique does not need any prior guess and it is sufficiently confident when using it in moderate reverberation [25]. Sources' TDOA's are determined from the cross-correlation function of successive frames from Hilbert Envelope (HE) of Linear Prediction (LP) residual all over the mixed speech. The occurred number of each delay is computed along the recorded mixture. The number of speakers is the number of superiors 'peaks' , and there TDOA's are determined by their locations with reference to the zero time lag. Knowing that each involved speech source has a specific TDOA, we can exploit this feature de design a specific weight function to each source.

3.1.1. Weight function design

For the design of weight function, we assume the sparseness property of source signals [14]. Based on this assumption, it is likely that at most only one source is predominant in each time-frequency observation.

A 12^{th} order LP analysis is performed on the two observed mixtures. Sources' excitation characteristics are extracted from Hilbert Envelop (HE) of the Linear Prediction residual (LPr) of the two observed mixtures, noting that using the HE of the LP residual of a signal allows more highlighted peaks. These peaks denote Instants of Excitations of Glottal Closure (GCI's) of different sources involved in the observed mixture. HE's of the LP residual are more preprocessed by dividing the square of each sample of the HE by the moving central average of the HE computed over a short window around the sample [25].

Estimating speech source signals is based essentially on separating their excitation peaks. This is done by aligning normalized preprocessed HE's of the LP residual of each

mixed speech captured by each microphone: HE's of LP residual $h_1(n)$ captured by Microphone1 is kept as reference, and the HE's of LP residual $h_2(n)$ captured by Microphone2 is shifted by the delay d_i of the i^{th} desired source. Then considering $h_{si}(n)$ the minimum of the sequence $h_1(n)$ and $h_2(n\text{-}d_i)$ as follows:

$$h_{si}(n) = min\ \left(h_1(n) - h_2(n - d_i)\right) \quad (4)$$

Where $i \in \{1\ 2\ 3\}$, h_{s1}, h_{s2}, h_{s3} are sequences retaining GCI's of S_1, S_2, and S_3, respectively.

Computing a specific weight function relies on exploiting GCI's of the desired source to relatively enhance it. In fact we need to emphasize GCI's of the desired source from GCI's of competing sources. It is performed by computing the difference between $h_{sj}(n)$ of the j^{th} desired source and $h_{si}(n)$ of the i^{th} undesired source where $i \neq j$.

In the following, we detail the case where we want to separate S_1 from a convolutive mixture involving three simultaneously speaking speakers ($S_1, S_2,$ and S_3).

We proceed by computing $h_{12}(n)$, and $h_{13}(n)$ as follows:

$$h_{12}(n) = h_{s1}(n) - h_{s2}(n) \quad (5)$$

Where h_{12} is the difference showing GCI's of S_1as positive peaks, and GCI's of undesired S_1 as negative ones. This can be repeated by emphasizing $S_1$from S_3 by computing h_{13}(n) as follows:

$$h_{13}(n) = h_{s1}(n) - h_{s3}(n) \quad (6)$$

A linear combination is computed to emphasize GCI's of S_1 relative to that of S_2, and S_3 as follows:

$$h_{p1}(n) = h_{12}(n) + h_{13}(n)/2 \quad (7)$$

Since sources are overlapping (all sources are simultaneously speaking), they are inherently sparse, which means that some regions specific to a source are less affected by competing sources. This inherent sparseness is exploited to enhance a desired source from competing ones by computing a weight function to enhance regions around GCI's of that source. Thus, detected GCI's, which are emphasized by the linear combination function, are exploited to compute an LP weight function for each speaker. It is derived at two different levels, namely gross and fine as it's defined in [22].

The gross weight (W_{gi}) function is derived to identify desired and undesired i^{th} speaker regions in a mixed signal. It's computed by smoothing and normalizing the absolute value of $h_{pi}(n)$ by 100 ms Hamming window then nonlinearly mapping the smoothed sequence by a sigmoidal nonlinear function.

The fine weight (W_{fi}) function is computed to identify the location of significant excitation of desired and undesired sources in a mixture. $h_{pi}(n)$ values are smoothed with a 2 (ms) Hamming window. GCI's locations of the desired speaker are detected by convolving the positive values with the first order Gaussian differentiator (FOGD) [28], and GCI's locations of the undesired sources are detected by convolving absolute of negative values with FOGD. These locations are smoothed by a 3ms hamming window and used to derive the fine weight function of the desired source S_i.

The gross and fine weight functions are combined by a simple multiplication (W_{ci}) and its sample values are used to synthesize the desired image source $\widehat{s_{ij}^{img}}$ from the convolutive mixture.

3.1.2. Sources' spatial image estimation

The mixed speech signal is segmented into frames of 40ms overlapped by 10ms. Each frame is weighted by a Hamming window then subjected to a Discrete Fourier Transform (DFT) termed X(k).

The pitch and harmonics indexes, termed l_i, are used to select the p_i indexes by examining each short spectrum of each frame $X(k)$ in the range $l_{i-2} < p_i < l_{i+2}$ to pick peaks in the spectrum frame nearest to the N_p harmonics.

A window function W(k) for sampling magnitude of pitch and harmonics of each frame is computed as follows:

$$W(k) = Conv\{P(k), h_r(k)\} \quad (8)$$

Where

$$P(k) = \sum_{i=1}^{N_p} \delta(k - p_i) \quad (9)$$

$$h_r = \begin{cases} 1\ , -2 < k < 2 \\ 0\ , otherwise \end{cases} \quad (10)$$

Each sampled spectrum speech frame is enhanced depending on the VUD and the combined weight function sample values W_c(k). Fig.1 details the separation algorithm steps for each mixed spectrum frame where $A_f = 2$ is a multiplication factor [27], and $\beta = 0.02$ is the spectral floor [28].

Frames, which are subjected to the separation algorithm, are used to synthesize estimated speech sources' image using Inverse Discrete Fourier Transform (IDFT) then Overlap and Add approach (OLA) [29].

4. EXPERIMENTAL EVALUATION

Experiments are performed on "test" live recorded datasets which are taken from the development data of the fifth community-based Signal Separation Evaluation Campaign (SiSEC 2015) [30]. The sampling frequency is 16 kHz. The time duration of all individual sources is 10s. We considered mixtures which four audio speech sources including unrelated male and female speech. We have taken just mixtures with $1m$ microphones spacing and under low reverberation time ($RT_{60} = 130ms$). We note that applying the TDOA estimator algorithm leads to the estimated TDOA values perfectly matched with the true ones provided by the dataset.

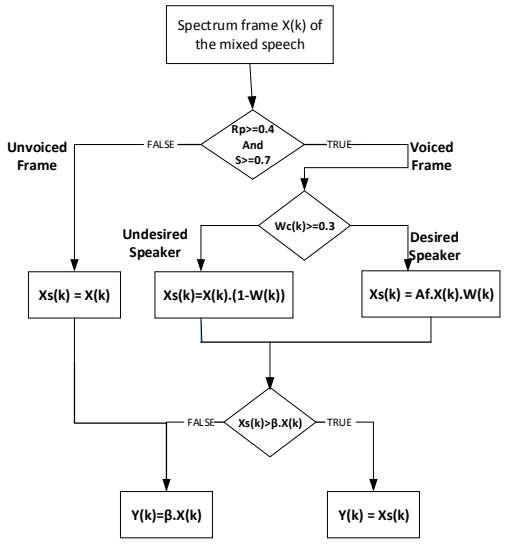

Fig.1 Detailed spectral enhancement diagrams: The estimation of desired speaker spectrum frame from the observed mixed one using its corresponding combined weight function values frame.

In order to evaluate the separation performance, commonly used objective metrics are used: BSS-EVAL toolkit [31], and PEASS toolkit [32]. The BSS-EVAL covers Signal-to-Distortion Ratio(SDR), Signal-to-Interference Ratio (SIR), Image-to-Spatial distortion Ratio (ISR) and Signal-to-Artifact Ratio (SAR) criteria expressed in decibels (dB), as defined in [31]. These criteria account respectively for overall distortion of the target source, residual crosstalk from other sources, spatial distortion and artifacts. The perceptual correspondence of these criteria are obtained from the PEASS toolkit which includes the Overall Perceptual Score (OPS), the Target-related Perceptual Score (TPS), the Interference-related Perceptual Score (IPS), and the Artifact-related Perceptual Score (APS) expressed in terms of a figure between 0 and 100. These criteria measure respectively how close the separated signal is to the clean reference signal, how well the target is preserved in the separated signal compared to the reference, how much interferences are cancelled, and the quality of the separated signal in terms of having no artifacts.

In order to evaluate our algorithm performance, we compare it with Nguyen's method performance which is posted in the SiSEC 2015 website and reported in [33]. Nguyen's method is similar to the Time-Frequency Masking approach [17] with a multi-band alignment permutation .

For a general overview, results are given based on the average over all sources and the observed mixtures, then listed in Table.1.

Nguyen's method introduces artifacts in the separated speech. This is due to the temporal fluctuations, which causes a relatively low SAR (SAR= 6.4 dB). However introduced artifacts are lower since it is expressed by a higher (SAR= 8.2 dB) when the Proposed Approach (PA) is used. This proves that PA is better than the Time-Frequency algorithm in rejecting artifacts.

Table 1 Separation results for the Proposed approach (PA) and Nguyen's method in terms of SDR, SIR, SAR, ISR measured in (dB) and in terms of OPS, TPS, IPS, APS measured in (%) for SiSEC 2015 Live recordings dataset with 1m microphone spacing and 130ms reverberation time. Mixtures are recorded by two microphones (Mic) and involving four speech sources (src).

	2Mic/4src			
Method	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS
PA	-5,7	1,6	-3,6	8,2
	8,4	55,1	1,0	83,3
Nguyen	4,5	8,3	8,0	6,4
	36,9	62,2	51,0	48,7

These findings correlate with performances in terms of APS. In fact, PA is still better in rejecting artifacts since it is expressed by a higher APS. By using PA, we reach ($APS = 83.3\%$), however we note a significant lower result performed by the Time–Frequency masking technique ($APS = 48.7\%$). Such improved artifact rejection performance obtained by PA is realized at the expense of significant introduced interference showed by low SIR. Similar results on trade-off between improvements in interference rejection (SIR) versus achieving a lower amount of artifacts (SAR) were noted when employing Nguyen's method. According to the SDR results, the Time-Frequency masking achieves statistically significant better performance in comparison to PA. In fact, as the separated signals were not time-aligned with respect to the original signals, the SDR and the SIR scores are negative. Although PA is outperforming in rejecting musical noise, it may be noted that its robustness in rejecting distortion and interferences strongly depends on the used TDOA estimator. In fact, as we have shown previously [25] estimated TDOA's are sensitive to reverberation and increasing number of speakers which make TDOA estimation ambiguous. Therefore, wrong weight function amplitude leads to ambiguous Voiced/Unvoiced decision. That's why, we assume that distortion is essentially due to Time Delay estimation error.

5. Conclusion

In this paper, a two-stage approach was introduced for separating multi-speech sources in a stereo convolutive mixture scenario. In the first stage, observed mixtures are framesized into Voiced/ Unvoiced frames and then the predominant source in each time frequency bin is identified through a specific weight function based on sources' excitation characteristics extraction. The weight function design relies on sources' counting and localization by determining sources' Time Delay of Arrival. It has been shown that this technique can enhance the quality of the estimated speech signal by evaluating it over convolutive live recorded mixtures using objective and perceptual metrics. Results have shown that the proposed technique is very efficient in rejecting artifacts. Nevertheless, this technique is still limited by reverberant conditions, where the used TDOA estimator algorithm becomes unavailable. Therefore, more improvement needs to be done in practical conditions.

6. REFERENCES

[1] T.-W. Lee, Independent Component Analysis—Theory and Applications.Norwell, MA: Kluwer, 1998.

[2] Unsupervised Adaptive Filtering (Volume I: Blind Source Separation), New York: Wiley, 2002.

[3] A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing: Learning algorithms. In Wiley Volume 1

[4] Jang, G.-J., and Lee, T.-W., & Oh, Y.-H. "Single-channel signal separation using time-domain basis functions". *IEEE Signal Processing Letters,* vol. 10(6), pp. 168–171, 2003.

[5] Araki, S., Mukai, R., Makino, S., Nishikawa, T., & Saruwatari, H. "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech". *IEEE Transactions on Speech and Audio Processing,* vol. 11(2), pp. 109–116, 2003.

[6] Asano, F., Ikeda, S., Ogawa, M., Asoh, H., & Kitawaki, N. "Combined approach of array processing and independent component analysis for blind separation of acoustic signals". *IEEE Transactions on Speech and Audio Processing*, vol. 11(3), pp. 204–215, 2003.

[7] Buchner, H., Aichner, R., and Kellermann, W. "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics". *IEEE Transactions on Speech and Audio Processing*, vol. 13(1), pp. 120–134, 2005.

[8] Smith, D., Lukasiak, J., and Burnett, I. "Blind speech separation using a joint model of speech production". *IEEE Signal Processing Letters*, vol. 12(11), pp. 784–787, 2005.

[9] Koldovsky, Z., and Tichavsky, P. Time-domain blind audio source separation using advanced ICA methods. *In Proc. interspeech*, Antwerp, Belgium, 2007, pp. 27–31.

[10] Das, N., Routray, A., & Dash, P. K. "ICA methods for blind source separation of instantaneous mixtures: a case study". *Neural Information Process*. *Letters and Reviews*, vol. 11(11), pp. 225–246, 2007.

[11] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," *in Proc. ICASSP 2000*, vol. 5, 2000, pp. 2985–2988.

[12] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[13] N. Roman, D.Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.,* vol. 114, no. 4, pp. 2236–2252, 2003.

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

[15] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments*," in Advances in Neural Information Processing Systems 19,* B. Sch¨olkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[16] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, pp. 382–394, Feb. 2010.

[17] H. Sawada, S. Araki, S. Makino, "A Two-Stage Frequency-Domain Blind Source Separation Method for Underdetermined Convolutive Mixtures," *in WASPAA 2007,* pp. 139-142, Oct. 2007.

[18] V. G. Reju, S. N. Koh and I. Y. Soon, "Underdetermined Convolutive Blind Source Separation via Time-Frequency Masking," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, NO. 1, Jan. 2010, pp. 101–116.

[19] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," *in Proc. WASPAA 2007*, 2007, pp. 147–150.11

[20] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures", *in Proc. WASPAA 2007*, Oct. 2007, pp. 139–142.

[21] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, pp. Article ID 24 717, 12 pages,2007.

[22] P. Krishnamoorthy, and S.R. Mahadeva Prasanna Two speaker speech separation by LP residual weighting and harmonics enhancement. Springer. *Int J Speech Technol* 13: 117–139, 2010.

[23] Proakis, J. G., & Manolakis, D. G. *Digital signal processing principles, algorithms, and applications* (3rd ed.). Upper Saddle River: Prentice Hall.1996.

[24] Markel, J. "The SIFT algorithm for fundamental frequency estimation". *IEEE Transactions on Audio and Electroacoustics*, vol. 20, pp. 367–377, 1972.

[25] M. Bouafif, and Z. Lachiri, "TDOA Estimation for Multiple Speakers in Underdetermined Case", *in Proc. INERSPEECH 2012*, vol 2, pp. 1746–1749, 2012.

[26] Kumara Swamy, R., Sri Rama Murty, K., & Yegnanarayana, B. "Determining number of speakers from multispeaker speech signals using excitation source information". *IEEE Signal Processing Letters*, vol. 14(7), pp. 481–484, 2007.

[27] Krishnamoorthy, P., & Prasanna, S. R. M. "Processing noisy speech by noise components subtraction and speech components enhancement". *In Proc. int. conf. systemics, cybernetics and informatics*, Hyberabad, India. 2007.

[28] Berouti, M., Schwartz, R., & Makhoul, J. "Enhancement of speech corrupted by acoustic noise". *In Proc. IEEE int. conf. acoust., speech, signal process*. pp. 208–211. 1979.

[29] J. Allen, L. Rabiner. "A unified approach to short- time Fourier analysis and synthesis". *Proc. IEEE*, vol. 65(11), pp.1558-1564, 1977.

[30] S. Araki, A. Ozerov, B.V. Gowreesunker, H. Sawada, F.J. Theis, G. Nolte, D. Lutter and N.Q.K. Duong, The 2010 Signal Separation Evaluation Campaign (SiSEC2010): Audio source separation *, in Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, pp. 114-122, 2010.

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

[32] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046-2057, 2011.

[33] Nobutaka Ono, Zafar Ra_i, Daichi Kitamura, Nobutaka Ito, Antoine Liutkus. The 2015 Signal Separation Evaluation Campaign. (LVA/ICA), Aug 2015, Liberec, France.