International Conference on Control, Engineering & Information Technology (CEIT'14)
Proceedings - Copyright IPCO-2014
ISSN 2356-5608

Harmonics Enhancement for Determined Blind Sources Separation using Source's Excitation Characteristics

Mariem Bouafif
LSTS-SIFI Laboratory
National Engineering School of Tunis
Tunis, Tunisia
mariem.bouafif@gmail.com

Zied Lachiri
Depart. Of Physic and Instrumentation
National Institute of Applied Sciences and Technology
Tunis, Tunisia
zied.lachiri@enit.rnu.tn

Abstract— **We present an improved method on combining temporal and spectral processing approaches for multichannel determined blind sources separation. The separation task is performed by applying the spectral processing on a mixed speech, using sources' excitation characteristics. The performance of the proposed method is investigated by separating two sources from a stereo recording mixture extracted from BSS-Locate [1]. Evaluation is performed by objective quality measure BSS-eval tool [2], perceptual evaluation of speech quality (PESQ), and Short-time Objective Intelligibility Measure (STOI) [3]. Simulations allow comparison with an existing spectral processing approach (TSP), and clearly demonstrate the efficiency and the outperformance of the proposed method.**

Keywords— Speech separation; LP residual; Glottal Closure Instants; time delay of arrival; Hilbert Envelop

I. INTRODUCTION

Extracting a target speech from a mixed stereo recording is one of the most important challenges in speech processing. In this field several approaches have been previously studied in the literature. Existing methods classified into three categories: The first approach exploits independent component analysis (ICA), called blind source separation (BSS) [4], [5], [6], [7], [8], [9], [10], and [11]. The second approach is the computational auditory scene analysis (CASA) [12], [12], [13], [14],[15], and [16].The third approach, called beamforming [17], is a type of spatial averaging which produces the greatest enhancement when the wanted components display significantly more inter-channel correlation than the unwanted components.

However, there are speech specific approaches (SSA) using speech specific features [18], [19], [20], [22], [23], [24], and [25]. The work presented here has focused on the improvement of the performance of an SSA technique by combining temporal and spectral processing.

The work by Krishnamoorthy and Prasanna [25] is based on applying a spectral processing technique on a temporally processed separated speech. This method is straightforward in low reverberant conditions. However, since the temporally processed speech is based on the use of an all-pole filter derived from the mixed speech, distortion still high in the estimated speaker's speech. The present study performs separation by applying the spectral processing on the mixed speech using temporal processing parameters. Comparing it with the TSP by Krishnamoorthy and Prasanna [25], the proposed method is more effective on term of separation and intelligibility. The conceptual block diagram of the existing TSP approach and the proposed one is shown in Fig.1.

The rest of the paper is organized as follows: The proposed method in the determined context is detailed in section 2. Experimental conditions, results, and various subjective measures are given in section 3. Finally, section 4 gives summary, conclusions and future scope of the present work.

II. THE PROPOSED APPROACH

The main field in the proposed approach is extracting a target speech source from a mixed one in the determined case, where we have two speakers speaking simultaneously and detected by two microphones, in low reverberant conditions.

The problem could be described using the Short-Time Fourier Transform (STFT).

$$X(t,f) = \sum_{n=1}^{2} d_n(f) S_n(t,f) + B(t,f) \qquad (1)$$

Where $X(t,f) = [X_1(t,f)\ X_2(t,f)]^T$ is the STFT of the observed signals at the two microphones, $S_n(t,f)$ is the n^{th} source signal in time frame t and frequency bin f, and d_n is the Time Delay of Arrival (TDOA) of the n^{th} source signal. The mixture $X(t,f)$ can be modeled as the sum of n delayed sources and reverberation $B(t,f)$.

The approach comprises two parts: temporal processing, and spectral processing. For this, we propose the use of the Hilbert envelope (HE) of the LP residual derived from the speech signal by linear prediction (LP) analysis [26], and [27].

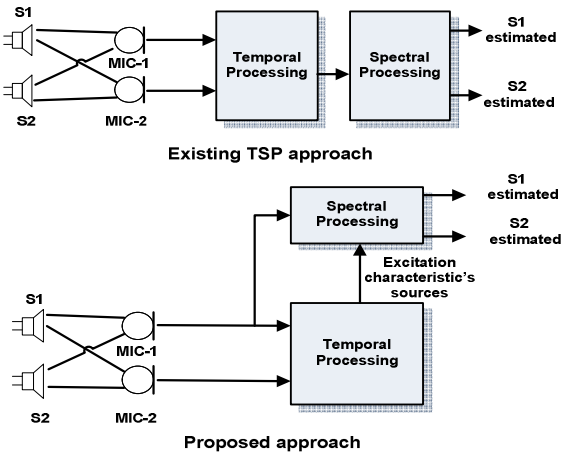

Fig. 1. Block diagram of the TSP approach [25], and the proposed approach.

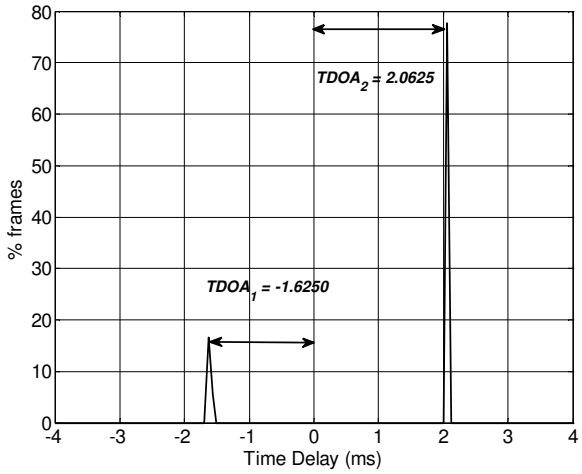

Fig. 2. Percentage of number of frames of each speaker as function of delays for a mixed speech of two speakers.

In the followed section a description of the proposed approach for two speaker's speech separation is detailed.

A. Temporal Processing

The temporal processing approach relies essentially on speaker's TDOA, GCI's detection of each source, and LP weighting.

1) Speaker's Time Delay of Arrival: The speaker's number, in a multi-sources mixed speech, as well as their different time delays, is determined using a method based on the excitation source components. This approach was already presented and evaluated in previous work [28]. The TDOA's are computed from the cross-correlation function of successive frames from HE's of LP residual (500ms shifted by 20ms) all over the mixed speech. The occurred number of each delay (in term of number of samples) is computed along the mixed speech. The number of speakers is the number of superiors 'peaks, and there TDOAs are determined by their locations with reference to zero time lag as shown in Fig 2.

2) *Source's Glottal Closure Instants Detection*:

The determination of GCI's from the speech signal is crucial. It's based on the HE's of LP residual of each observed mixed speech detected by the two sensors. HE's of the LP residual are 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 [29]. The normalized preprocessed HE's of the LP residual h_1(n) and h_2(n) of each mixed speech captured by each microphone are aligned after compensating the delay d_i of the i^{th} desired speaker. Competing speaker instants are in incoherence, whereas instants of the desired speaker are in coherence. By considering $h_{si}(n)$ the minimum of the sequence h_1(n) and h_2(n-d_i), only the instants referring to the i^{th} desired speaker are retained.

The difference between the HE's h_{s1} and h_{s2} is computed as follows:

$$h_{12}(n) = h_{s1}(n) - h_{s2}(n) \quad (2)$$

$$h_{21}(n) = h_{s2}(n) - h_{s1}(n) \quad (3)$$

Where h_{12} is the difference showing the instants of significant excitation of Spk1 as positive peaks, and the instants of significant excitation of Spk2 as negative ones, and vice versa for h_{21}.

3) LP weighting function: Enhancing desired speaker from competing one is performed by computing an LP residual weight function for each speaker derived at two different levels, namely gross and fine levels as it's defined in [25].

The gross weight function is derived to identify desired and undesired speakers regions. It's computed by smoothing and normalizing the absolute value of the separated HE's by 100 ms hamming window, then nonlinearly mapping the smoothed sequence by sigmoidal nonlinear function.

A fine weight function is then computed to identify the location of significant excitation of desired and undesired speaker (GCI's) in a mixed speech. First, the difference values of the separated preprocessed HE's are smoothed with a 2 ms hamming window. Then, GCI's locations of the desired speaker are detected by convolving the positive values with the first order Gaussian differentiator (FOGD) [30]. Whereas, GCI's

locations of the undesired speaker are detected by convolving absolute of negative values with FOGD. The fine weight function is derived by convolving the detected instants with a 3ms hamming window.

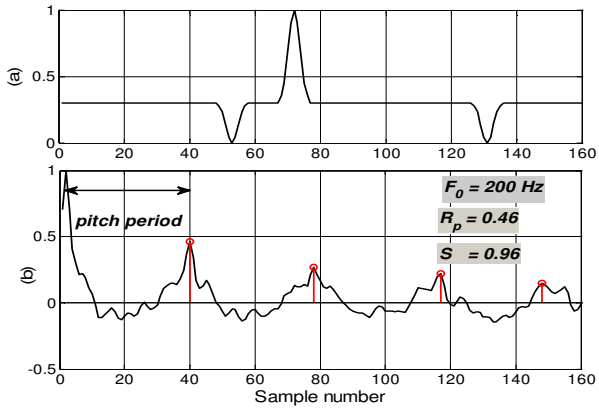

Fig.3. (a) Fine weight function frame specific to speaker1. (b) Normalized autocorrelation R (l) of mean subtracted HE of temporally weighted LP residual of the corresponding voiced frame mixed speech sampled at 8 kHz (tow speakers speaking simultaneously).

The LP residual of the observed mixed speech is weighted by the combined function W_c, computed by multiplying the gross and the fine weight functions, and then used to excite a time varying all-pole filter to synthesis the temporally estimated speech of the desired speaker.

4) Spectral Processing

As the desired spectrum could be reconstructed by using the separated harmonics, pitch detection and voiced unvoiced decision of each speaker's speech are crucial in spectral processing.

1) Pitch estimation: In this work, the pitch estimation is obtained from the normalized autocorrelation of the mean subtracted HE's of the LP residual of the mixed speech [31]. It is frame-sized in blocks of 40ms overlapped by 10ms, and then subjected to a normalized autocorrelation [32].

As the minimum possible frequency F_0 of a human speech is 50 Hz, we seek the correlation sequence over the lag range [-20ms: 20ms]. Then we take the half of the autocorrelation of each block, as it's just mirror for real signal. As the maximum human pitch F_0 is 500 kHz, we search for the first major peak with reference to zero time lag between 2ms (500kHz) and 20ms (50 kHz) [33].

2) Voiced Unvoiced decision: The voicing decision is made by computing the magnitude of the first major peak R_p [34], and similarity behaviour S [31]. Each frame of speech, subjected to autocorrelation, is considered as voiced only if $R_p \geq 0.4$ [34], and $S \geq 0.7$ [31]. It can be observed from Fig.3 (a) that the fine weight function enhances GCI's of the desired speaker and deemphasize GCI's of undesired speaker. The pitch is obtained in this voiced frame will be of the desired speaker as shown in Fig 3 (b).

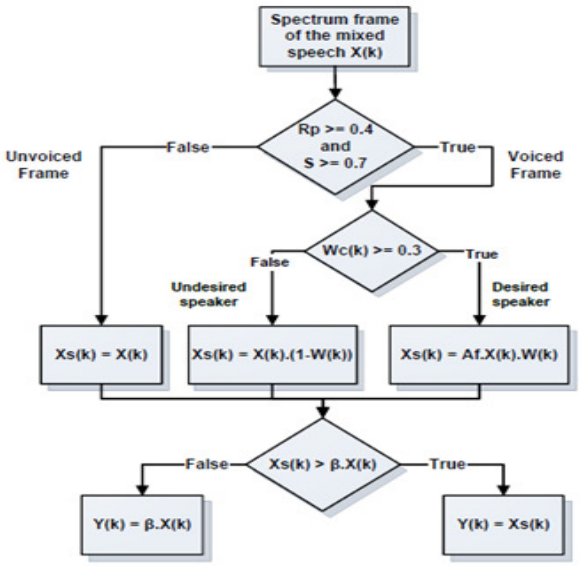

Fig.. 4. Detailed spectral processing diagram to enhance desired speaker spectrum frame from the observed mixed one using its corresponding combined weight function values frame.

3) Speaker's speech estimation: First, the degraded 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 $X(k)$. Second, 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.

The third step is to compute the window function $W(k)$ for sampling magnitude of pitch and harmonics of each frame as follows:

$$W(k) = Conv\{P(k), h_r(k)\} \quad (4)$$

Where

$$P(k) = \sum_{i=1}^{N_p} \delta(k - p_i) \quad (5)$$

$$h_r = \begin{cases} 1, -2 < k < 2 \\ 0, otherwise \end{cases} \quad (6)$$

Each sampled spectrum speech frame is enhanced depending on the voiced unvoiced decision and the combined weight function sample values $W_c(k)$, as it is explained in Fig.4 where $A_f = 2$ is a multiplication factor [35], and $\beta = 0.02$ is the spectral floor [36]. The separated signal is synthesized using Inverse Discrete Fourier Transform (IDFT) then Overlap and Add approach (OLA) [37].

Table 1: OBJECTIVE MEASUREMENT PERFORMANCE, AVERAGED OVER THE TWO SPEAKERS IN DIFFERENT MIXTURE EXTRACTED FROM BSS-LOCATE TOOLBOX [38] ,ACHIEVED BY TEMPORAL SPECTRAL PROCESSING APPROACH (TSP) [25] COMPARED TO THE PROPOSED APPROACH (PA) ON TERM OF SDR IMPROVEMENT (DB), SIR IMPROVEMENT (DB), PESQ, AND STOI. AVG: IS THE AVERAGE OF EACH METRIC OVER ALL MIXTURES. THE BEST RESULT IN EACH METRIC IS HIGHLIGHTED IN BOLD FACE.

	TSP				PA			
	SDR_imp	SIR_imp	STOI	PESQ	SDR_imp	SIR_imp	STOI	PESQ
Mix1	0,41	5,55	0,68	2,08	4,90	5,54	0,82	2,54
Mix2	-0,17	1,45	0,59	1,46	4,39	7,41	0,74	2,05
Mix3	-3,45	2,18	0,54	1,34	1,41	2,96	0,69	1,88
Mix4	-2,05	2,39	0,65	1,93	2,10	2,70	0,78	2,42
Mix5	-3,12	1,68	0,63	1,78	1,15	1,88	0,76	2,25
Mix6	-1,73	3,18	0,66	1,94	2,81	3,60	0,79	2,45
Mix7	-3,06	4,20	0,55	1,20	1,87	4,51	0,70	1,92
Avg	-1,88	2,95	0,61	1,68	**2,66**	**4,09**	**0,76**	**2,22**

III. EXPERIMENTAL DATABASE AND EVALUATION METRICS

The proposed approach and TSP [25] algorithms were coded in Matlab. We performed experiments to separate two speech sources captured by two microphones. We considered the same mixture signals as in [1], which are available as part of the BSS Locate toolbox [38]. We had different mixed speech containing two sources (male and female) in different configurations, at 50 ms reverberation time, and sampled at 16 kHz.

Separation performance of approaches was evaluated with respect to the signal-to-distortion ratio (SDR), and signal-to-interference ratio (SIR) criteria expressed in decibels (dB), as defined in [39]. These criteria account respectively for overall distortion of the target source, and residual crosstalk from other sources. The separation performance was evaluated in terms of SDR and SIR improvements, as it is defined in [40], and we took the average over two speakers. To evaluate the intelligibility of the estimated sources, we also conducted an objective test on term of Perceptual Evaluation Speech Quality (PESQ) [41], and the Short-time Objective Intelligibility Measure (STOI) [3].

IV. RESULTS AND DISCUSSION

This subsection is devoted to compare the potential source separation performance achievable by the proposed approach with TSP proposed by Krishnamoorthy and Prasanna [25].

The resulting source separation performance in terms of SDR, SIR improvement, PESQ, and STOI is depicted in Table.1. Interestingly, the proposed approach outperforms TSP in term of SDR improvement, and SIR improvement, over all mixtures. TSP shows poor distortion rejection performance. As it's suspected, the distortion still high in the separated speaker's speech performed over all mixtures.

It's due to the all-pole filter derived from the mixed speech used to synthesize the temporally processed speech. Such low distortion rejection performance explains the moderate intelligibility of the separated speech (STOI = 0.61). In fact, the difference in speech intelligibility performance between the two approaches is significant. It reaches 0.82, for the first mixture, whereas it's only equal to 0.68 performed by the TSP approach.

The proposed approach provides an average improvement in the perception quality performance of 32% compared to the TSP approach. It reaches 2.54 for the first mixture, however it's only equal to 2.08 performed by TSP approach.

V. CONCLUSIONS

We presented a novel algorithm for blind source separation, based on the temporal and the spectral approaches. The combination of these two methods exists in previous work, known as TSP. It applies the spectral processing on the temporally processed speech. In our work, we tried to improve this combination. We applied the spectral processing approach on the mixed speech using sources' excitation characteristics of the temporally processed speech. Results show that our method outperforms TSP in term of intelligibility and separation.

Even if the proposed approach outperforms TSP, it still limited by the reverberation. Our proposed method is based on the Time delay of arrival estimation over a linear prediction residual approach, which fails in underdetermined high reverberant environment [18].In future work, we will try to improve the proposed approach by employing a more robust TDOA estimator, and we will try to extend it to the underdetermined context.

REFERENCES

[1] C. Blandin, A. Ozerov and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering", Signal Processing, vol. 92, pp. 1950–1960, August 2012.

[2] E. Vincent, R. Gribonval, and C. Fevotte. "Performance measurement in blind audio source separation". IEEE Transactions on Audio, Speech and Language Processing, vol. 14 (4), pp. 1462–1469, jul 2006.

[3] C.H.Taal, R.C.Hendriks, R.Heusdens, J.Jensen "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech", ICASSP 2010, Texas, Dallas.

[4] Jang, G.-J., and Lee, T.-W. "A maximum likelihood approach to single-channel source separation". Journal of Machine Learning Research, vol. 4, pp. 1365–1392. Special issue on independent components analysis, 2003.

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

[6] 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.

[7] Asano, F., Ikeda, S., Ogawa, M., Asoh, H., and 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.

[8] 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.

[9] 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.

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

[11] Das, N., Routray, A., and 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.

[12] Brown, G. J., and Cooke, M. "Computational auditory scene analysis. Computer Speech and Language", vol. 8(4), pp. 297–336, 1994.

[13] Wang, D ,and Brown, G. J. "Computational auditory scene analysis: principles, algorithms, and applications". New York: Wiley- IEEE Press, 2006, pp. 395.

[14] Slaney, M. The history and future of CASA. In Divenyi, P. (Ed.) Speech separation by humans and machines pp. 199–211. Norwell: Kluwer Academic, 2005.

[15] Brown, G. J., and Wang, D. Separation of speech by computational auditory scene analysis. In Benesty, J., Makino, S., and Chen, J. (Eds.) Speech enhancement (pp. 371–402). Berlin: Springer, 2005.

[16] Radfar, M. H., Dansereau, R. M., and Sayadiyan, "A. Monaural speech segregation based on fusion of source-driven with model driven techniques". Speech Communication, vol. 49(6), pp. 464–476, 2007.

[17] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K. "Blind source separation combining independent component analysis and beamforming''. EURASIP Journal of Applied Signal Processing, vol. 11, pp. 1135–1146, 2003.

[18] Parsons, T. W "Separation of speech from interfering speech by means of harmonic selection". The Journal of the Acoustical Society of America, vol. 60, pp. 911–918, 1976.

[19] Hanson, B., and Wong, D. "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech". In Proc. IEEE int. conf. acoust., speech, signal process. vol. 9, pp. 65–68, 1984.

[20] Lee, C. K., and Childers, D. G. "Cochannel speech separation". The Journal of the Acoustical Society of America, vol. 83, pp. 274–280, 1988.

[21] Quatieri, T. F., and Danisewicz, R. G. "An approach to cochannel talker interference suppression using a sinusoidal model for speech". IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.38, pp. 56–69, 1990.

[22] Morgan, D. P., George, E. B., Lee, L. T., and Kay, S. M. "Cochannel speaker separation by harmonic enhancement and suppression". IEEE Transactions on Speech and Audio Processing, vol. 5, pp. 407– 424, 1997.

[23] Yegnanarayana, B., Prasanna, S. R. M., and Mathew, M. "Enhancement of speech in multispeaker environment". In Proc. European conf. speech process, technology, Geneva, Switzerland pp. 581–584, 2003.

[24] Mahgoub, Y. A., & Dansereau, R. M. "Time domain method for precise estimation of sinusoidal model parameters of co-channel speech". Research Letters in Signal Processing. doi:10.1155/2008/364674, 2008.

[25] P. Krishnamoorthy, and S.R. Mahadeva Prasanna "Two speaker speech separation by LP residual weighting and harmonics enhancement". Springer Science+Business Media, LLC 2010. Int J Speech Technol, 2010.

[26] J.Makhoul: "Linear prediction: A tutorial review". Proc. IEEE vol. 63 pp. 561, 580, 1975.

[27] Ananthapadmanabha, T. V., and Yegnanarayana, B. "Epoch extraction from linear prediction residual for identification of closed glottis interval". IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, pp. 309–319, 1979.

[28] M. Bouafif, and Z. Lachiri, "TDOA Estimation for Multiple Speakers in Underdetermined Case", in Proc. 13th Ann. Conf. of Int speech Comm Asso 2012 (INERSPEECH 2012), vol 2, pp. 1746–1749, 2012.

[29] Kumara Swamy, R., Sri Rama Murty, K., and 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.

[30] Prasanna, S. R.M., and Subramanian, A. "Finding pitch markers using first order Gaussian differentiator". In Proc. IEEE third int. conf. intelligent sensing information process, Bangalore, India, vol. 1, pp.140-145.

[31] Prasanna, S. R.M., and Yegnanarayana, B. "Extraction of pitch in adverse conditions". In Proc. IEEE int. conf. acoust, speech, signal process, Montreal, Quebec, Canada vol. 1, pp. I-109–I-112, 2004.

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

[33] Naotoshi ,Seo sonots, "ENEE632 Project 4 Part I: Pitch Detection." March 24, 2008

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

[35] Krishnamoorthy, P., and 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.

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

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

[38] [Online] available: http://bass-db.gforge.inria.fr/bss_locate/.

[39] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in Proc. Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA), pp. 552–559, 2007.

[40] S. Araki, H. Sawada, R. Mukai and S. Makino, Underdetermined Blind Sparse Source Separation for Arbitrarily Arranged Multiple Sensors, Signal Processing, vol.87, pp. 1833- 1847, August 2007.

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process, vol. 16, no. 1, pp. 229–238, Jan. 2008.