Bouafif Mansali, Mariem; Bäckström, Tom; Lachiri, Zied

Evaluation of Zero Frequency Filtering based Method for Multi-pitch Streaming of Concurrent Speech Signals

Published in:
28th European Signal Processing Conference, EUSIPCO 2020 - Proceedings

DOI:
10.23919/Eusipco47968.2020.9287322

Published: 24/01/2021

Document Version
Publisher's PDF, also known as Version of record

Evaluation of Zero Frequency Filtering based Method for Multi-pitch Streaming of Concurrent Speech Signals

Mariem Bouafif Mansali
El Manar University, SITI laboratory, National Engineering School of Tunis
Tunis, Tunisia
mariem.bouafif@gmail.com

Tom Bäckström
Aalto University, Department of Signal Processing and Acoustics
Espoo, Finland
tom.backstrom@aalto.fi

Zied Lachiri
El Manar University, SITI laboratory, National Engineering School of Tunis
Tunis, Tunisia
zied.lachiri@enit.utm.tn

***Abstract*—Multiple pitch streaming from a mixture is a challenging problem for signal processing and especially for speech separation. In this paper, we use a Zero frequency filtering (ZFF) based new system to stream pitch of multiple concurrent speakers. We propose a workflow to estimate pitch values of all sources in each single frame then streaming them into trajectories, each corresponding to a distinct source. The method consists of detecting and localizing the involved speakers in a mixture, followed by a ZFF based approach where involved speakers' pitches are iteratively streamed from the observed mixture. The robustness of the proposed system is tested over two, and three overlapping speech mixtures collected in reverberant environment. The results indicate that our proposal brings ZFF to a competitive level with another recently proposed streaming approach.**

Keywords—Pitch estimation, Zero Frequency Filtering, Epochs, Multipitch, Streaming

I. INTRODUCTION

Multi-pitch analysis is the task of analyzing, detecting and separating the pitch frequency contours of multiple speakers from their mixture. The difference between the pitch frequency contours of speakers is generally the most important feature used by the human auditory system to separate speakers. Accordingly, multi-pitch analysis has become a major field of study as it plays an important role in different issues, especially in automatic music transcription [1], source separation [2], melody extraction [3], multi-talker speech recognition [4] prosody analysis [5], and cocktail party problem [6]. The multi-pitch analysis problem has been broken down into three different levels [7]: (1) the multi-pitch estimation level is to estimate pitch values of all interfering sources without determining their sources. (2) The multi-pitch tracking level is to connect pitch estimates in adjacent frames to form continuous pitch trajectories that typically correspond to individual notes or syllables. (3) The multi-pitch streaming level is to stream pitch estimates into different pitch trajectories over the entire mixture involving unvoiced, and silence discontinuities. Multi-pitch analysis has given a rise for a wide variety of methods. Most of them perform at the estimation level, [8]–[10]. Particularly for speech, the multi-speech context has been addressed in different works [11]–[14]. Many other algorithms address the tracking level [15]. However, few studies have investigated the streaming level. In these studies, Bayesian network approaches [16]–[18] were especially designed for polyphonic signal. Also a few attempts were made for the multiple speakers' case. In this category, a supervised method is proposed based on factorial hidden Markov [19]. Recently, an unsupervised method [20] has been developed and used for the separation task. Moreover, the latest general and unsupervised constrained-clustering approach was proposed by Duan, Han, and Pardo [21]. It explores different timbre features for music and speech. This method does not require pre-training, However, it requires a prior knowledge of the number of sources.

The objective of this study is to examine the feasibility of using a Zero frequency filter (ZFF) based system to address the streaming level. Since ZFF was developed for the analysis of speech, it has been widely employed for both epoch extraction [22]–[23] and pitch estimation [24]. Recently, Yegnanarayana and Prasanna [25] briefly proposed a discriminative technique to separate speaker excitation features in a two-speaker mixture, then use it to generate a pitch track to each interfering source showing the case of one and two speakers. They expected that the proposed approach will be accurate even for more than two speakers. A ZFF based technique for voiced/unvoiced (V/UV) discrimination has also been investigated by Yegnanarayana and Prasanna [26], where they reported performance improvement under noisy conditions. These two works motivate us to study the feasibility of using ZFF for multi-talkers streaming system which, unlike Yegnanarayana and Prasanna tracking algorithm, streams estimated pitches over the entire conversation. We extended the tracking algorithm for a variable number of speakers, then we incorporate a voice activity detector to reach the streaming level. This step-up should allow our new approach to cope robustly with long utterances containing both voiced and unvoiced portions.

Our contributions are as follows. First, we develop a ZFF based method for streaming pitches of multiple simultaneously speaking speakers. Second, we evaluate the proposed streaming approach under varying conditions. Finally, we investigate the competitivity of our approach by comparing its performance with one of the latest multi-pitch streaming systems using new state-of-the-art metrics.

II. PROPOSED ZFF BASED METHOD FOR MULTIPITCH STREAMING

In this section, we outline the main steps required for a ZFF based streaming approach. We recall the empirical procedure for a conventional ZFF. Then, we detail the different components of our multi-pitch streaming system.

A. Method overview

In our previous work [27], we proposed a blind speech separation technique based on epochs detection and segregation. When evaluated on the Signal Separation Evaluation Campaign [28], our preliminary system was able to be improved unless we enhance the epochs extractor. Since our preliminary system of [27], our work has focused on building a robust epoch and pitch estimator. In this work, we take one step forward and propose to apply a ZFF method inside the epoch detector in order to extract discriminative epochs at frame level. The algorithm presented in this paper performs multi-pitch streaming using the segregated detected epochs of each source.

As summarized in the flowchart illustrated in Fig.1, the proposed system can be divided into four main steps: 1) pre-processing of the observed mixtures to enhance epochs of involved speakers. 2) The number of involved speakers and their Time Delay of Arrival (TDOA) are estimated using a localization algorithm applied on the observed mixtures. 3) epochs segregation and pitch estimation at frame level are performed using ZFF from both the original and the time delayed pre-processed mixtures based on the estimated TDOA in step2. 4) Final trajectory formation is cast by smoothing estimated pitches in every detected voiced frame all over the entire conversation, including voiced/unvoiced decisions. An iterative programming is applied from steps 2 and 3 to arrive at a final stream to each involved source.

The proposed algorithm is frame based. It involves several experimentally determined parameters such as frame duration, frame shift, voicing decision threshold, window size. These parameters are reported in this paper along with values used for experimental.

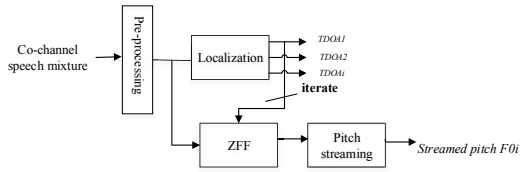

Fig. 1. The block diagram of the proposed multi-pitch streaming approach. A cochannel input mixture is processed, to localize involved speakers, followed by the iterative feature extraction then pitch streaming of each source referring to its TDOA.

B. Conventional ZFF

The algorithmic steps of conventional ZFF method [23] are summarized as follows:

A speech signal $s[n]$ is pre-emphasized using a difference operation. Then the pre-emphasized speech signal $x[n]$ is passed through two ideal digital resonators called zero-frequency resonator (ZFR) whose center frequency are located at 0 Hz. The resulting output signal $y_1[n]$ is subjected to trend removal by a local mean subtraction as follows:

$$y_2[n] = y_1[n] - \frac{1}{2M+1}\sum_{p=-M}^{+M} y_1[n+p], \qquad (1)$$

where 2M + 1 is the size of the window corresponding to the average pitch period computed over a longer segment of speech. The trend removed signal $y_2[n]$ is referred as the zero Frequency Filtered signal (*ZFFs*) where each positive peak is hypothesized as an epoch location [29].

C. Iterative multi-pitch streaming

a) Pre-processing: Pre-processing consists of creating a new version of observed mixtures to enhance the excitation components then facilitate their extraction. Each observed mixture is independently processed by using multiple functions. Since experimental evaluations indicated better epochs extraction by using the Hilbert envelop (HE) of the Linear prediction (LP) residual of a speech signal, rather than applying it on the signal itself particularly in the multi-talkers context [23]–[30], the HE of LP residual is applied on both observed mixtures. A Gabor filter on LP residual was also used for the epochs' strength increasing [31]. Applying a Gabor filter on the HE of LP residual of the signal will reduce significantly the effects of all formants. In the following, we refer to the filtered HE of the LP residual as the pre-processed signal.

b) Localization: One of the key features of the proposed approach is the use of localization information to guide pitch streaming. In this paper, we consider the multi-speakers scenario where speakers are independent, and their voice activities are unknown to the system. To stream their pitches, it is necessary to separate the excitation information corresponding to each speaker. As each source has its own localization, its TDOA is specific to that source. Hence, referring to speakers' TDOA, the excitation information corresponding to each speaker could be separated. For that, a TDOA approach was developed and tested in a preliminary work where we exploited the point property of the impulse-like excitation to localize speakers in a co-channel mixture [32]. In fact, the number of speakers and their TDOAs are synthesized over a cross-correlation function between the modified and pre-processed HE of LP residual of the two observed mixtures. The TDOA MATLAB code is available [33].

c) Epochs extraction using ZFF: One way of determining a pitch of a speaker is to first locate its excitation components, then separate them from other competing sources in the mixture. The following procedure was used to extract excitation components of speech candidates in a mixture.

- The excitation components corresponding to a specific target source are enhanced by a channel time delay alignment which makes the desired source's excitation components in coherence, whereas the competing speakers' excitation components will be incoherent. To emphasize excitation components of the desired speaker, and de-emphasize those of the competing speakers, we consider the minimum function $m_i[n]$ from the two pre-processed signals. Therefore, relatively high values in the minimum function indicate excitation regions of the i^{th} target speaker speech. Some biased epochs which belongs to other speakers are further reduced by considering a subtraction function $f_{ij}[n]$, enhancing i^{th} source's excitation impulses from j^{th} ones for each pair of candidates as follows:

$$f_{ij}[n] = m_i[n] - \alpha\, m_j[n], \qquad (2)$$

where $i = [1, 2 \ldots I]$, $j = [1,2 \ldots I]$ and $j \neq i$. Here, I is the total estimated number of speech sources. $\alpha = 0.001$ is a small fraction.

- To emphasize excitation impulses of S_i relative to that of all other talking sources, we use a linear combination $P_i[n]$ computed as follows:

$$P_i[n] = \frac{1}{(I-1)} \sum_{j=1}^{I \neq j} f_{ij}[n], \quad (3)$$

where $i = [1, 2 \ldots I]$, $j = [1, 2 \ldots I]$, and $j \neq i$.

- Finally, the i^{th} speaker's epochs are derived by applying ZFF on $P_i[n]$ as described in the subsection B. This procedure is iteratively repeated to determine i^{th} speaker's epochs sequences until the estimated number of sources in the mixture is reached. Fig. 2 (a) plots a pre-processed speech segment and Fig. 2 (b) plots its $ZFFs$ in each sample point. All detected epochs belong to the unique active source.

The result of this process is a sequence of epochs for each involved candidate over the entire mixture.

d) Pitch streaming: The primary use of the detected epochs associated to each candidate for each frame is the instantaneous pitch estimation. It is simply identified by calculating the reciprocal of the time interval between adjacent successive epochs' locations. As a speech signal can involve unvoiced/silence discontinuities, such regions should be identified to stream the detected pitch all over the entire signal. As the regions of glottal activity are more significant in voiced regions than in unvoiced ones, detected epochs can be used as an aid for making voicing decisions by referring to their peak intensity, known as the strength of excitation (SoE). It is defined as the slope of $ZFFs$ at each epoch [29], and it is measured by computing the difference between the negative and positive sample values on either side of each detected epoch. The V/UV clustering of speech is done by fixing a threshold on SoE. Typical threshold is experimentally found to be 50% of the average SoE value of the entire speech utterance. The V/UV regions clustering is further validated based on the pitch period and the instantaneous jitter measured at each epoch to avoid any spurious epoch that could occur in the silence or unvoiced regions. Only epochs with pitch period less than $15ms$ and a jitter within $1ms$, are considered as voiced. The SoE is shown by Fig.2. (c) where epochs hypothesized as voiced, obtained after pitch period and jitter based validation, are under a dashed line.

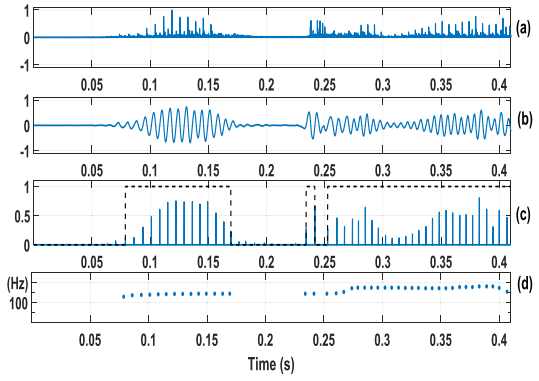

Fig. 2. Computation of pitch using ZFF for a speech segment (a) the pre-processed signal. (b) Zero-frequency filtered signal (ZFFs) of (a). (c) strength of excitation (SoE) of impulse calculated from the ZFFs. Voiced regions are marked by dashed line (d) Pitch detection in voiced regions using ZFF applied on HE of LP residual of the speech signal.

After V/UV decisions, all instantaneous pitch values are simply identified then concatenated and viewed as one continuous pitch stream in the detected voiced regions as shown in Fig. 2. (d). The variation of the number of epochs, their strength of excitation and the interval between successive epochs during voiced regions vary with the used window size (frame length) in the mean subtraction (1) [28]. For the present work, a window size of $(2M + 1)$ is the average pitch period which we update for every short time segment of $30ms$ shifted by $20ms$. We note that used frame length is typical of this used in many speech processing applications.

III. Experimental Evaluation

A. Database description

One of the greatest difficulties in co-channel multi-speech pitch streaming evaluation is the lack of high-quality pitch referenced data set. In the field of speech separation, there are large stereo recorded data set with multi-speech content, making possible the localization and source counting task for the proposed approach. For the evaluation task, we use multiple mixtures extracted from BSS-Locate toolbox [34]. Selected mixtures contain two, or three sources of the same gender (male, or female), and recorded under varying reverberation times $RT_{60} = [50, 150, 250, 500](ms)$. Optimal channel distance was set to 1m to insure accurate TDOA results [32].

Streaming evaluation requires reference pitch streams. However, used mixtures are without provided ground truth pitch values. Therefore, we generate our own auto-labeled pitch values of the original clean speech, which are subsequently used for the performance evaluation as ground truth pitch values. For this, well performing algorithms have been employed: *BaNa* [35], *HPS* [36], *Praat* [37], *Cepstrum* [38], and *Pefac* [39]. These algorithms have been evaluated on auto-labelled databases with available ground truth pitch values and score about 85% of accuracy. Accordingly, an accurate ground pitch value has been approximated to the average value of pitch estimated over the cited algorithms. A frame is assumed as voiced, and the detected pitch value is considered as ground value if the estimated pitch delivered from all algorithms is within a frequency difference threshold set to 10%. Otherwise, the frame is considered as unvoiced with no pitch ground value.

B. Error measures

To measure progress, streamed speakers' pitches are evaluated using five metrics [11]: (1) The transition error rate E_{xy} is defined as the percentage of time frames where x pitch points are misclassified as y pitch points. It is also called deletion error when the number of pitches in each frame is less than the true number of reference pitches in each frame $(x < y)$. However, when the estimated pitches are greater than the number of the reference pitches $(x > y)$, the transition error are defined as insertion error. (2) The gross error rate E_{Gross} is the percentage of frames where the deviation between estimated and reference pitch is larger than 20% in Hz. (3)The fine error rate E_{Fine} is the percentage of frames where the estimated pitch deviation from the reference pitch is smaller than 20% in Hz. This frequency threshold is commenly used for pitch analysis of speech. E_Gross, and E_Fine are determined for all involved speakers. (4) The overall error noted as E_{Total} is the sum of all error terms.(5) The accuracy is defined and used in [21]. A bijection between

the i ground-truth pitch trajectories and the i estimated trajectories is made by choosing the assignment leading to the best overall multi-pitch streaming accuracy.

C. Experimental results

a) The effect of pre-processing: To evaluate the effect of pre-processing, we computed the accuracies, the gross errors, and the fine errors for three conditions by applying ZFF on: the observed unprocessed co-channel signal, the HE of LP residual signal, and the pre-processed signal. We perform the simplest case where we have only two concurrent speakers. Fig.3 reports accuracies, gross errors and fine errors averaged over ten mixtures. It shows that even if fine errors are very similar for all input signal versions, the pre-processing is beneficial in terms of gross error. Consequently, significant improvement of accuracy (with more than 50%) is performed by pre-processing the observed mixture. These results, show that pre-processing the signal plays an important role in improving the performance of the algorithm.

b) The effect of reverberation: It could be questioned whether or not the ZFF based approach is adequate in the presence of realistic reverberation. For that, we study the effect of varying reverberation time conditions on the performed accuracy in two-speaker case. Results are reported in Fig. 4 where we note that unsurprisingly, the proposed algorithm is subject to large decreases in accuracy as reverberation increases. It is essentially due to insertion errors increasing with reverberation which affects epochs detection accuracy by inserting extra false peaks.

c) Comparison with other multi-pitch streaming algorithm: Another performance evaluation is conducted by comparing the proposed algorithm with a single-channel constrained clustering algorithm referenced as Duan's approach.We used Duan's code available on [35] and his suggested parameters settings. The code is applied on a randomized channel. Experiments are conducted for the case of two and three sources. Table. I reports detailed error rates averaged over ten mixtures recorded under low reverberation time $RT_{60} = 50\ (ms)$. We can note that the number of insertion errors in the case of two-speaker and three-speaker mixtures is significantly high compared to deletion error rates. A preliminary study revealed that these errors are essentially due to biased TDOA estimation affecting sources' epochs segregation. Moreover, there are regions in speech where one speaker is dominated by the other making incorrect pitch detection. We can also remark that the most influencing insertion errors are E_{12}, E_{13}, and E_{23} compared to E_{01}, E_{02}, and E_{03}, which ensure that our algorithm yields the gross error from insertion errors in voiced frames. Even if that error is considerably high, V/UV regions detection is properly done and it can be moreover ensured by the small deletion errors rate E_{10}, E_{20}, and E_{30}.

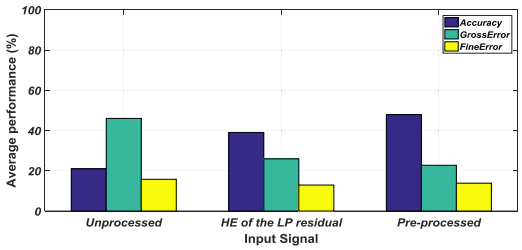

Fig. 3. The effect of pre-processing the input signal on the performance of ZFF for multi-pitch streaming.

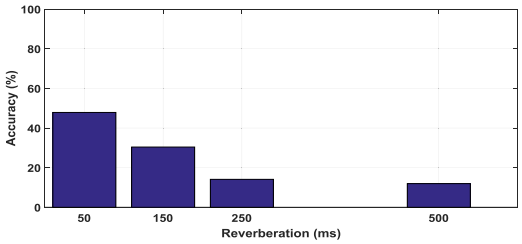

Fig. 4. Performance accuracy (%) at various reverberation times RT_{60}(ms)

It is worth to note that the delay between the source and its image at each channel may leading to a delay between computed ground-truth pitch trajectory and the estimated pitch trajectory as they are extracted respectively from provided clean source and its observed image detected in a mixture. That delay due to the room impulse response may affect the resulting accuracy. We add that very close results are noticed all overs mixtures for every type of error. We mention that having an overall error rate exceeding 100% may occur when frames are subject to both insertion (or deletion) and gross errors together. We can also obviously note that contrary to the suggested approach, Duan's algorithm performance shows 0% of deletion errors, however a considerably high insertion errors E_{13}, and E_{23} are noticed leading to a remarkable gross error. These two errors rate are also the main contributors to E_Total. We note that using the same gender of speakers makes pitch trajectories interfere with each other, and many must-link constraints imposed by Duan's algorithm by consequence are incorrect and lead subsequently to a large insertion error. Even if we outperform Duan's algorithm, it is worth to note that 47% of accuracy still not satisfying. Introducing additional pre-processing stages may be possible way to overcome the reverberation effect. Such issue is quit challenging for the Duan's algorithm even at low reverberation.

IV. Conclusion

In this paper, a ZFF based multi-pitch streaming algorithm has been developed which combines multiple processing steps to segregate a pitch stream from multiple-speaker co-channel mixture. Although similar ZFF based methods have

TABLE 1 RESULTS FOR THE SUGGESTED APPROACH ON *BSS* DATABASE. THE TRANSITION ERROR RATE E_{xy} IS THE PERCENTAGE OF TIME FRAMES WHERE x PITCH POINTS ARE MISCLASSIFIED AS y PITCH POINTS. ALL ERROR MEASURES ARE IN PERCENTAGE (%)

Mixtures	Algorithm	Insertion Errors						Deletion Errors						E_Gross	E_Fine	E_Total	Accuracy
		E_01	E_02	E_03	E_12	E_13	E_23	E_10	E_20	E_30	E_21	E_31	E_32				
2 speakers	Proposed	3,7	3,17	-	37,04	-	-	1,06	1,06	-	5,82	-	-	22,75	13,85	88,46	**47,87**
	Duan's	0	7,94	-	56,08	-	-	0	0	-	0	-	-	5,29	4,38	73,69	38,7
3 speakers	Proposed	0,53	2,12	2,12	9,52	9,52	37,57	0	0,53	0	05.29	1,06	1,59	68,25	13,5	150,5	**38,15**
	Duan's	0	0	4,23	0	23,28	57,14	0	0	0	0	0	0	13,23	6,41	104,3	23,18

been used to some epoch detection, pitch tracking and V/UV decision algorithms for a known number of speakers, these methods have been improved, implemented and integrated in the current proposed algorithm to create a multi-pitch streamer applicable for an unknown number of speakers. An analysis of errors confirmed that the proposed approach compares favorably with another multi-pitch streamer, especially for low reverberated mixtures. Nevertheless, additional improvements need to be done for a more accurate results making our system more suitable for some technological uses.

REFERENCES

[1] A. Klapuri and M. Davy, "Signal Processing Methods for Music Transcription," Springer, 2006.

[2] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music souce separation by average harmonic structure modeling," *IEEE Trans. Audio Speech Language Processing*, vol. 16, no. 4, pp. 766–778, 2008.

[3] J. Han and C.-W. Chen, "Improving melody extraction using probabilistic latent component analysis," The 36th International Conference on Acoustics, Speech, and Signal Processing, May 22–27, Prague Congress Center, Proceedings, pp. 33–36, 2011.

[4] M. Cooke, J. R. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, pp. 1–15, 2010.

[5] D.-n. Jiang, W. Zhang, L.-q. Shen, and L.-h. Cai, "Prosody analysis and modeling for emotional speech synthesis," in *The 30th International Conference on Acoustics, Speech, and Signal Processing, March 18-23, Philadelphia, PA, USA*, *Proceedings,* pp. 281– 284, 2005.

[6] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal of the Acoustic Society of America*, vol. 25, pp. 975–979, 1953.

[7] https://www.music-ir.org/mirex/wiki/MIREX_HOME

[8] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[9] A. de Cheveigne and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, pp. 175–185, 1999.

[10] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *Journal of the Acoustical Society of America*, vol. 119, pp. 2498–2517, 2006.

[11] M. Wu, D. Wang, and G. Brown "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process*., vol. 11, no. 3, pp. 229-241, 2003.

[12] F. Sha and L. Saul, "Real-time pitch determination of one or more voices by nonnegative matrix factorization," in 17th International Conference on Neural Information Processing Systems, December 13-18, Vancouver, British Columbia, Canada, proceedings pp. 1233–1240, 2004.

[13] F. Bach and M. Jordan, "Discriminative training of hidden Markov models for multiple pitch tracking," in The 30th International Conference on Acoustics, Speech, and Signal Processing, March 18-23, Philadelphia, PA, USA, Proceedings,2005, pp. 489–492.

[14] Z. Jin and D. Wang, "Hmm-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio Speech Language Processing*, vol. 19, no. 5, pp. 1091–1102, 2011.

[15] Z. Duan, J. Han, and B. Pardo, "Song-level multi-pitch tracking by heavily constrained clustering," in - The 35th International Conference on Acoustics, Speech, and Signal Processing, March 14 – 19, Dallas, Texas, USA, proceedings, pp. 57–60, 2010.

[16] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication,* pp. 337–349, 1999.

[17] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. Audio Speech Language Processing,* vol. 14, no. 1, pp. 91–98, 2006.

[18] M. Bay, A. F. Ehmann, J. W. Beauchamp, P. Smaragdis, and J. S. Downie, "Second fiddle is important too: pitch tracking individual voices in polyphonic music," *in The 13th International Conference on Music Information Retrieval, October 8-12, Porto, Portugal, proceedings,* pp. 319–324, 2012.

[19] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Trans. Audio Speech Language Processing*, vol. 19, no. 4, pp. 799– 810, 2011.

[20] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio Speech Language Processing*, vol. 21, no. 1, pp. 122–131, 2013.

[21] Z. Duan, J. Han, and B. Pardo, " Multi-pitch Streaming of Harmonic Sound Mixtures*" IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, pp. 138 – 150, 2014.

[22] B. Yegnanarayana, S. R. M. Prasanna and S. Guruprasad "Study of robustness of zero frequency resonator method for extraction of fundamental frequency" in ICASSP 2011 - *The 36th International Conference on Acoustics, Speech, and Signal Processing, May 22-27, Prague, Czech Republic, proceedings, pp.* 5392-5395, 2011.

[23] P. Gangamohan and B. Yegnanarayana, "A Robust and Alternative Approach to Zero Frequency Filtering Method for Epoch Extraction", in INTERSPEECH 2017 - *18th Annual Conference of the International Speech Communication Association, august 20-24, Stockholm, Sweden, Proceedings,* pp. 2297-2300, 2017.

[24] D. Pravena and D. Govind, "Expressive Speech Analysis for Epoch Extraction Using Zero Frequency Filtering Approach", in *IEEE Students' Technology Symposium,* 2016

[25] B. Yegnanarayana1 and S. R. Mahadeva Prasanna, "Analysis of instantaneous F0 contours from two speakers mixed signal using zero frequency filtering", in ICASSP 2010 *The 35th International Conference on Acoustics, Speech, and Signal Processing*, *March14 – 19, Dallas, Texas, USA, Proceedings*, 2010, pp. 5074-5077.

[26] N. Dhananjaya and B. Yegnanarayana, "Voiced/Nonvoiced Detection Based on Robustness of Voiced Epochs'' *IEEE Signal Processing Letters, vol. 17, no. 3, March 2010*

[27] M. Bouafif, Z. Lachiri "A new time-frequency approach for underdetermined convolutive blind speech separation, In ICASSP 2016 *The 41th International Conference on Acoustics, Speech, and Signal Processing,March 21-25, Shanghai, China, Proceedings, 2016, pp. 3226-3230.*

[28] Antoine Liutkus, Fabian Robert-Stöter, Zafar Rafii, Daichi Kitamura, Bertrand Rivet, et al.. The 2016 Signal Separation Evaluation Campaign. 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017), Feb 2017, Grenoble, France. pp.323-332.

[29] B. Yegnanarayana, K S R. Murty, S. Rajendran, "Analysis of stop consonants in indian languages using excitation source information in speech signal," in Workshop *Speech Anal. Process. Knowledge Discovery*, June 4–6 Aalborg, Denmark, proceedings 2008.

[30] M. Bouafif, Z. Lachiri, "Underdetermined blind source separation technique based on speech features extraction," *.I. J. Speech Technology issue 19, vol. 4,* pp. 697-706, 2016.

[31] S. Dixon, "On the computer recognition of solo piano music*,"* in ACMC - *Australasian Computer Music Conference, Brisbane, proceedings* 2000.

[32] M. Bouafif, Z. Lachiri "TDOA Estimation for Multiple Speakers in Underdetermined Case" INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012, pp. 1748-1751.

[33] M. Bouafif, Z. Lachiri "SRC_Num_TDOA: Multiple speech sources' number and their TDOA Estimation from a stereo recorded mixture," *SoftwareX,* Vol.5, pp. 234-242, 2016.

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

[35] http://www.ece.rochester.edu/projects.

[36] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental frequency measurement," Journal of the Acoustical Society of America, vol. 43, pp. 829–834, 196

[37] http://www.fon.hum.uva.nl/praat/.

[38] L. R. Rabiner and R. W. Schafer, "Theory and Application of Digital Speech Processing". *Pearson*, 2011.

[39] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.