Research Article

Improved comb filter for the separation of the voiced speech signals in the case of two speakers

Jihen Zeremdini[1] · Mohamed Anouar Ben Messaoud[1] · Aicha Bouzid[1]

© Springer Nature Switzerland AG 2019

Abstract

In this paper, we present a method for separating voiced sounds from a composite signal. This method is mainly based on the separation by modified comb filter. This filter is keyed to the average values of the estimated pitch. This estimation is performed through an autocorrelation of multi-scale product analysis to separate the effects of the source and the vocal tract. The "autocorrelation of the multi-scale product" method allows noise elimination and the apparition of a signal periodic structure. Peaks that appear are used to calculate the mean fundamental frequency of the target speaker which will be used in the corresponding comb filters to determine the target speaker contribution. After the subtraction of this contribution from the mixture, we obtain the intrusion speaker. This separation is validated by its application on Cooke database and a part of VCTK database and compared to recent methods as Wang–Brown, Hu–Wang, Zhang–Liu and Quan systems. Results confirm the performance of the proposed approach.

Keywords Autocorrelation of multi-scale product · Fundamental frequency f_0 · Improved comb filtering · Multi-speakers · Voiced sounds separation

1 Introduction

One of the most important issues in the field of speech processing concerns the estimation of the pitch in the presence of a range of interferences that are added to the main signal. We can cite some cases such as an environmental noise, reverberation or just another person speaking at the same time with the target speaker. During the last decades, various methods of separating sounds have been proposed, such as the blind source separation (BSS) which allows the extraction of the unknown speech signals from the mixture signals with no a priori information about the mixed signals and the sources [1–6]. We mention also spectral subtraction [7], subspace analysis [8], hidden Markov modeling [9] and sinusoidal modeling [10]. These approaches expect certain interference properties, and then, they segregate composite speech based on these hypotheses. These methods assume certain signal properties, but they have some limitations in comparison with the separation of human sounds.

Other approaches have been proposed as the CASA (computational auditory scene analysis) model [11–15] which are based on the principles of ASA "auditory scene analysis" [16] described by Bregman for separation of composite signals. Bregman supposed that the brain organizes the data according to the general properties of the signal (harmonicity, synchronization, etc.) by using automatic and native mechanisms. The low-level processing, data driven, allows the grouping of the signal components, thus providing auditory stream corresponding to external events.

A CASA system generally follows four steps: peripheral analysis, the extraction of sound properties, segmentation and grouping. The peripheral processing decomposes

✉ Jihen Zeremdini, zeremdini_jihen@hotmail.fr; Mohamed Anouar Ben Messaoud, anouar.benmessaoud@yahoo.fr; Aicha Bouzid, bouzidacha@yahoo.fr | [1]LR11ES17 Signal, Image and Information Technology Laboratory, National School of Engineers of Tunis, University of Tunis El Manar, 1002 Tunis, Tunisia.

SN Applied Sciences (2019) 1:818 | https://doi.org/10.1007/s42452-019-0840-6

Received: 25 February 2019 / Accepted: 26 June 2019 / Published online: 2 July 2019

the auditory scene representation in a two-dimensional time–frequency (TF) through a filter bandwidth and a windowing in function of time. The second step extracts auditory properties corresponding to the principles of ASA, which will be used later in the segmentation and grouping. In the segmentation and grouping, the system produces segments for concurrent voice and target voice (interference) and then groups the segments from the target in a stream. This flow corresponds to a sound source. Finally, the separate target signal will be synthesized from this flow [13].

In the different methods of speech separation, the pitch is a major parameter. Indeed, several multi-pitch search algorithms have been proposed, like in the case of the automatic transcription of polyphonic music [17] or in the case of multi-pitch estimation in composite speech. The first systems typically have had significant limitations in the content and were only capable of detecting two harmonic overtones [18, 19]. Among the multi-pitch estimation methods, we find mainly those based on statistical approaches such as Bayesian modeling [20] or the hidden Markov model (HMM) [21]. These techniques have been validated only in the case of two speakers or a single speaker with noise.

In this work, we use the comb filters in the frequency domain for the separation of composite speech signals. The comb filters have been used in [22, 23] on the synthesized signals for the separation of polyphonic sounds. But their use on real signals will be more complicated because of certain intrinsic phenomena to the production of voiced sounds such as the effects of noise due to the vocal tract, the variability of the frequency of the glottal excitation called jitter or the pitch interaction with the first forming.

Then, our principal goal is the target signal estimation in voiced zones.

This paper is organized as follows. Section 2 describes related works. Section 3 presents the principle of the proposed method for voiced sounds separation. The results and the comparison are given in Sect. 4. Finally, Sect. 5 concludes this work.

2 Related works

Multiple comb filters and autocorrelation of the multi-scale product for multi-pitch estimation approach presents a new method "that estimates the fundamental frequency in the case of a real noisy environment when many persons speak at the same time and considers the case of two speakers" [24].

On the other hand, the method of the evaluation of multi-pitch estimation by multi-scale product analysis describes "a multi-scale product comparative study for dominant pitch estimation in a mixture" [25].

Besides, multi-pitch estimation based on multi-scale product analysis, improved comb filter and dynamic programming approach is about "a method consists on the autocorrelation function of the multi-scale product calculation of the mixture signal, its filtered version by a rectangular improved comb filter and the dynamic programming of the residual signal spectral density for multi-pitch detection" [26].

Finally, the comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation work presents and compares "several studies that have used computational auditory scene analysis (CASA) for speech separation and recognition" [27].

Nowadays, researchers are increasingly interested in making systems capable of separating signals. In our case, we are focused on developing a method of voiced sounds separation from a composite signal which is mainly based on the use of a modified comb filter that is keyed to the average values of the estimated pitch. In addition, this estimation is performed through an autocorrelation of multi-scale product analysis to separate the effects of the source and the vocal tract.

So, our actual work is about the target signal estimation in voiced zones using the fundamental frequency estimation of the dominant signal. Absolutely, we use some techniques presented in previous works such as the autocorrelation, the multi-scale product analysis and comb filtering. But, we have modified the structure of the filter and used these tools for the separation after the pitch detection.

3 Description of voiced sounds separation system

For the separation of composite speech, we propose an approach based on the "autocorrelation of multi-scale product (AMP)" and the modified comb filtering (MCF). This filtering is applied only in the voiced regions. For the conception of our comb filter, we need to determine the true value of the fundamental frequency f_0 of the voiced regions of composite signal.

Then, the use of comb filters is dependent on pitch estimation of dominant speaker. To do this, we use a method based on the "AMP analysis." The proposed algorithm operates as shown in the block diagram given in Fig. 1.

3.1 Pitch estimation by the "AMP"

To decompose the composite signal into frames, we use a hamming window with 64-ms duration.

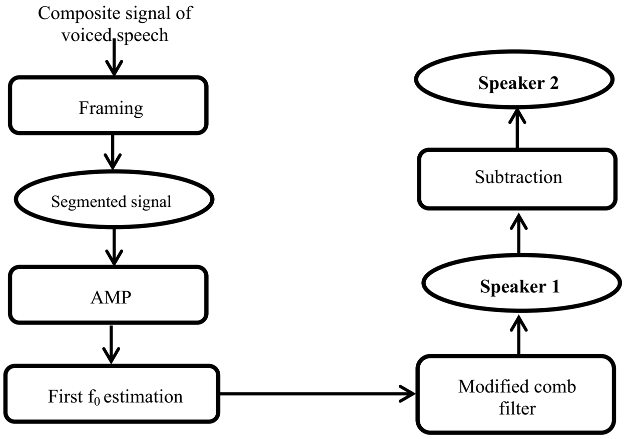

Fig. 1 Flow diagram of separation approach

In order to obtain the "multi-scale product (MP)" signal, we use a continuous wavelet transform (WT) at three successive scales (2^{-1}, 2^0 and 2^1) and we calculate their product of the coefficients.

In fact, the wavelet transform allows the characterization and the detection of signal singularities [24]. The wavelet used in this MP analysis is the quadratic spline function with a support of 0.8 ms. It is the first-order derivative of the scaling function, which closely approximates a Gaussian function.

The MP detects singularities present in the signal and allows filtering the signals and generating simplest periodic structures. Then, we calculate the autocorrelation function, which allows the estimation of the pitch of the target speech signal [28].

The "AMP" for the time interval i is described as follows:

$$\mathrm{AMP}_i(k) = \sum_{l=0}^{N-1} P_i(l)P_i(+k) \tag{1}$$

where N is the length of the analyzed frame and P_i is the multi-scale product.

On the other hand, for a signal x, the MP at scales s_j is given by:

$$P(n) = \prod_{j=1}^{J} Wx\left(n, s_j\right) \tag{2}$$

where $Wx\ (n, s_j)$ is the WT of $x(n)$ at the scale $s_j = 2^j$.

3.2 Improved comb filter and speakers separation

Comb filters are based on the summation of two identical signals together with a small delay between them. These filters allow selected frequencies to pass while blocking all other frequencies.

For an input signal $x[n]$, this process is described by the following equation:

$$Y[n] = x[n] + g * x[n - D] \tag{3}$$

where g is a factor, which scales the gain of the filter between $1 + g$ and $1 - g$, and D is the delay in samples.

There are two types of comb filters:

- a comb filter with finite impulse response (FIR)
- a comb filter with infinite impulse response (IIR)

These filters are described, respectively, by the following equations:

$$Y[n] = X[n] + \alpha X\left[n - k\right] \tag{4}$$

$$Y[n] = X[n] + \alpha Y\left[n - k\right] \tag{5}$$

where k represents the delay associated with the period T and α is a fixed parameter applied to the delayed signal.

The transfer functions associated with these filters are, respectively:

$$H(z) = 1 + \alpha z^{-k} \tag{6}$$

$$H(z) = \frac{1}{1 - \alpha z^{-k}} \tag{7}$$

Comb filters are used in numerous applications such as music analysis, speech enhancement and separation and multi-pitch estimation. In this context, a least mean square (LMS) adaptive filtering approach has been proposed for removing the deleterious effects of additive noise on the speech signal by Sambur [29]. Moreover, Darlington and Campbell [30] developed a multi-channel and sub-band adaptive system for enhancement of speech signals corrupted by background noise. We mention also Iliev and Kasabov technique [31] which is based on adaptive filtering with averaging (AFA) used for echo cancelation. Besides, we cite Christensen and Jakobsson method [32] which introduced a number of new filter designs that are constructed for periodic signals and that give full parameterizations of these signals.

In this work, to obtain the dominant signals, a rectangular comb filter is used. Indeed, we have used the comb filter in the frequency domain with a rectangular frequency response because the bands forming this type of filter are needed to filter properly the fundamental frequency and its harmonics.

This filter is tuned by the estimated f_0, and it has adaptable amplitude which varies according to the maximum amplitude of the mixture signal at each frame.

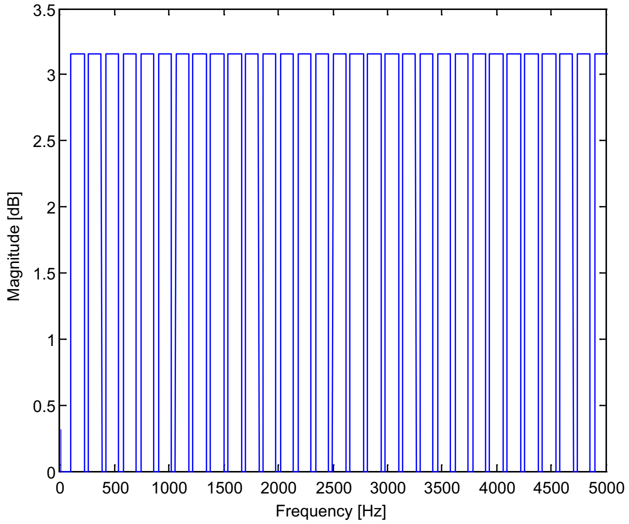

Fig. 2 Frequency response of the rectangular comb filter

The resulting signal from this filtering presents the first speaker or the dominant signal. To obtain the second speaker or the intrusion, we apply a subtraction between the mixture signal and the signal estimated by the designed comb filter.

Figure 2 shows the frequency response of the designed comb filter for a fundamental frequency $f_0 = 160$ Hz and a sampling frequency $Fs = 16$ kHz.

4 Experiments and comparison

4.1 Experiments

The separation of composite sounds is to keep only the harmonics of the signal and reject the intermediate frequency components. For separation of our composite signal, the proposed approach uses a comb filter with a rectangular bandwidth. The filter amplifies all multiple frequencies $f_0 = 1/T_0$ according to the width of the band. f_0 is given by the "AMP analysis."

This approach has been evaluated using Cooke database [33], which contains 100 mixture sounds obtained by mixing ten male voiced speech signals with ten other signals representing a variety of sounds. Interferences can be classified into three categories:

- Interferences without pitch formed by white and impulse noises.
- Interferences having a pitch quality formed by pure frequency of 1 kHz, cocktail party noise, rock music, siren and ring tone.
- Speech interferences, which are speech signal uttered by a woman 1, speech signal uttered by a man 2 and speech signal uttered by a woman 2.

The text pronounced by male target signals is "Why are you all weary." On the other hand, the text pronounced by the speech interference is "Don't ask me to carry an oily rag like that."

Signals constituting the Cooke database are monophonic sound reproduction and are sampled at the frequency 16 kHz.

The mixture signal is decomposed into frames using a hamming window with 64-ms duration.

Figures 3, 4 and 5 present the mixture, the reference target and the estimated target by our approach, respectively, with a noise from the first and the second categories of interferences (white noise, pure frequency of 1 kHz and cocktail party noise).

Then, for this type of mixture signals, we obtain good results for the estimated signals comparing with references signals.

Figures 6, 7 and 8 show the composite signal followed by the first speaker and the second speaker signals references and estimated from the separation operation using a rectangular comb filter.

Figure 6 describes the case when the mixture contains two male voices. This is the most difficult case in the separation task because the two speakers are from the same gender and have very similar characteristics like the fundamental frequency. This explicates why the estimated intrusion is not very similar to the reference one.

Then, we note that the estimated signal is approximately similar to the reference signal for both speakers

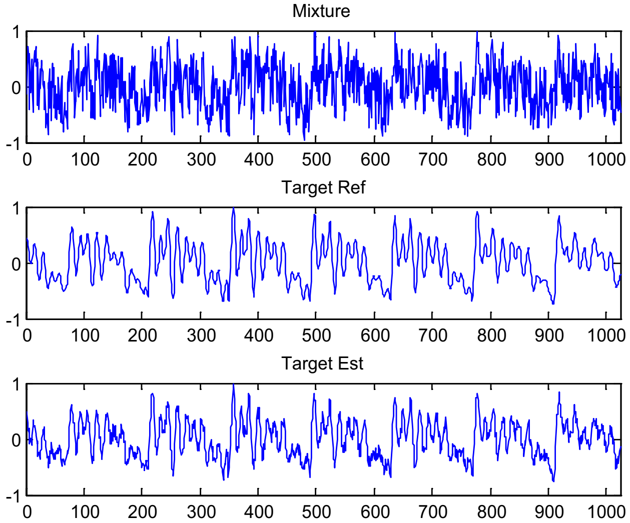

Fig. 3 Separated signals from composite speech from Cooke database (male voice and white noise)

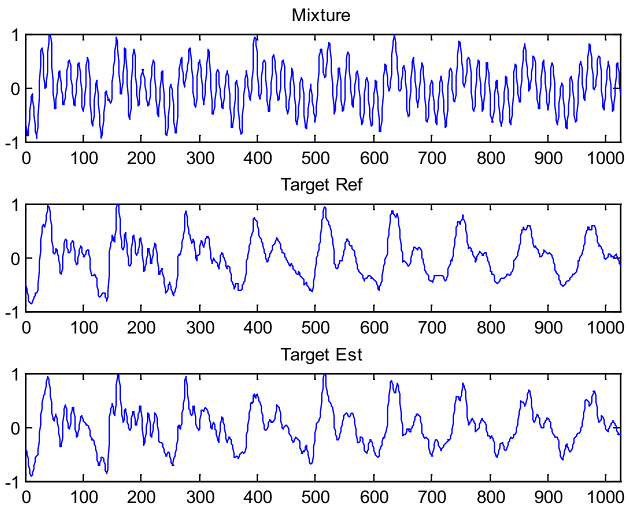

Fig. 4 Separated signals from composite speech from Cooke database (male voice and pure frequency of kHz)

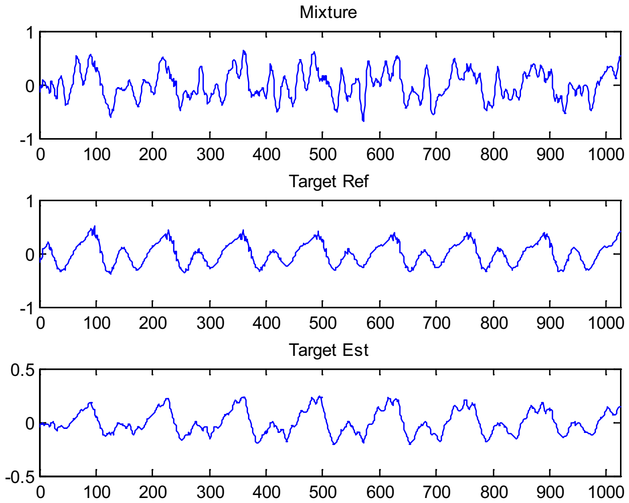

Fig. 5 Separated signals from composite speech from Cooke database (male voice and cocktail party noise)

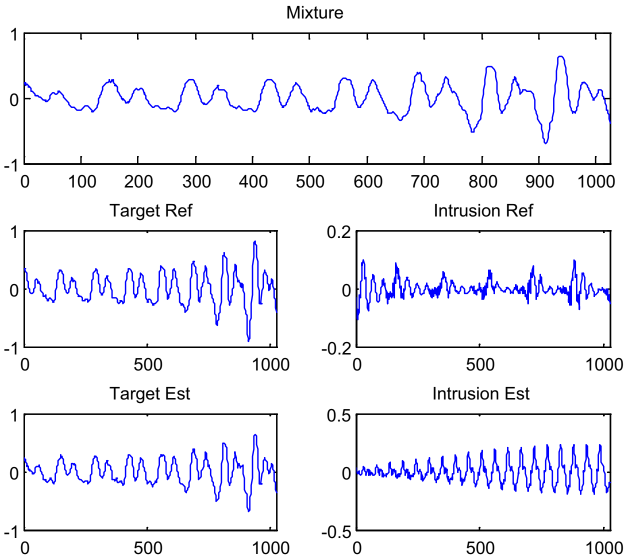

Fig. 6 Mixture and both reference and estimated target and intrusion frames of male voices (from the Cooke database)

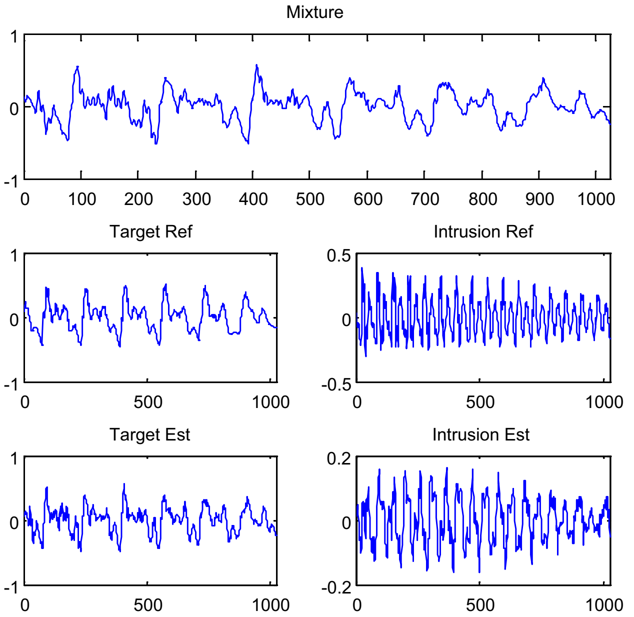

Fig. 7 Mixture and both reference and estimated target and intrusion frames of male and female voices (from the Cooke database)

composed the mixture signal in the three presented cases, especially for the mixture formed by male and female voices.

On the other hand, the proposed approach has been evaluated also in the VCTK database.

This database includes speech data uttered by 109 native speakers of English with various accents. Each speaker reads out about 400 sentences, most of which were selected from a newspaper plus the Rainbow Passage and an elicitation paragraph [34].

We have only used ten speakers from this database: two mixtures formed by two males, two mixtures formed by two females and six mixtures formed by one male and one female [35].

Figures 9, 10 and 11 present the mixture, the reference target, the estimated target by our approach and the estimated target by Quan et al.'s system, respectively, in the case of a mixture formed by two males, in the case of a mixture formed by two females and in the case of a mixture formed by one male and one female.

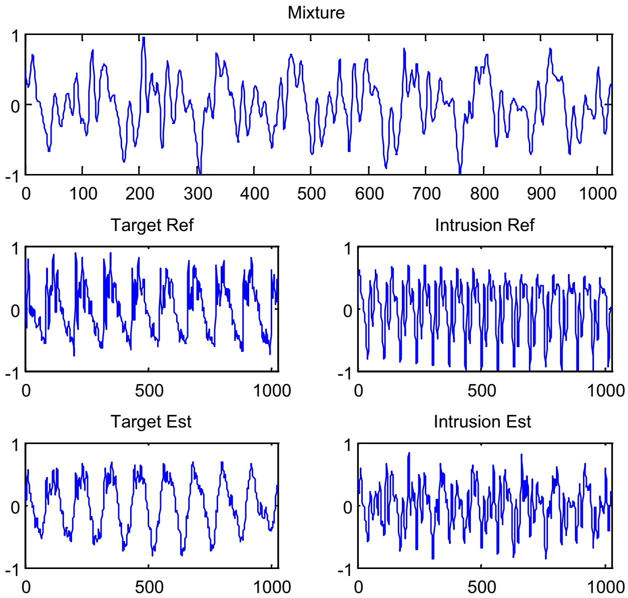

Fig. 8 Mixture and both reference and estimated target and intrusion frames of male and female voices (from the Cooke database)

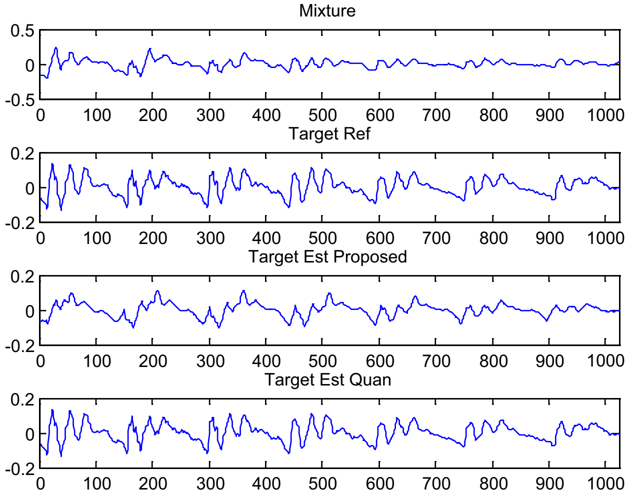

Fig. 9 Mixture, reference and both estimated target frames of two males voices (from the VCTK database)

From these figures, we can conclude that the estimated target signal by our proposed approach is approximately similar to the reference target signal in the three presented cases.

These results show the robustness of our proposed method.

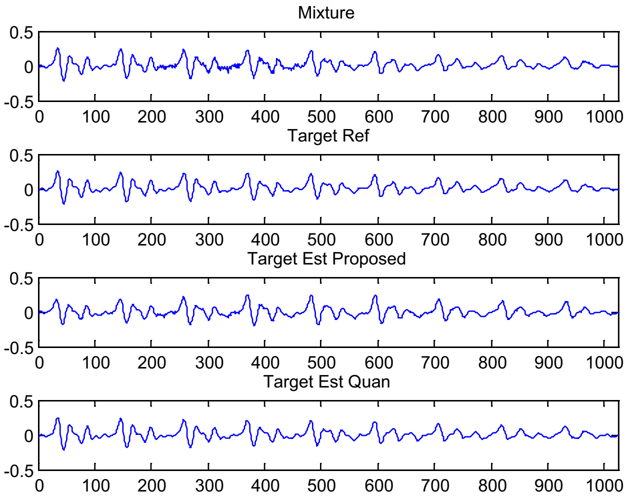

Fig. 10 Mixture, reference and both estimated target frames of two females voices (from the VCTK database)

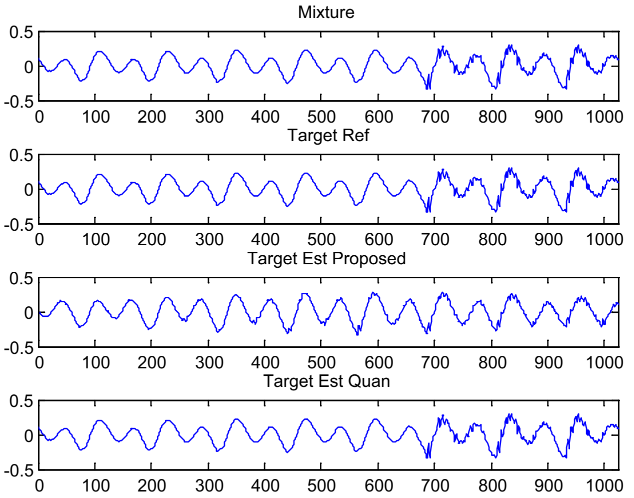

Fig. 11 Mixture, reference and both estimated target frames of one male and one female voices (from the VCTK database)

4.2 Comparison and discussion

Our approach performed well for both speakers estimation. It is compared to Wang and Brown [36], Hu and Wang [37] and Zhang and Liu [38] systems for speech separation.

Wang–Brown model [36] is based on a framework oscillatory correlation which is supported by recent neurobiological findings. F_0 information is available in the pooled correlogram. Indeed, after the formation of the correlogram by computing a running autocorrelation of the auditory nerve activity in each filter channel, these correlograms are computed at 10-ms intervals, forming a three-dimensional volume in which time, channel center frequency and autocorrelation lag are represented on orthogonal axes. Additionally, a "pooled" correlogram is formed at each time frame by summing the periodicity information in the correlogram over frequency. The largest peak in the pooled function occurs at the period of the dominant fundamental frequency (f_0).

On the other hand, Hu and Wang [37] proposed a tandem algorithm that performs the pitch estimation of the target signal and the segregation of voiced portions of target speech jointly and iteratively. They used the ideal binary mask (IBM) at the voiced frames of the target utterance to estimate a pitch period at each frame.

In addition, Zhang–Liu approach [38] concerns monaural voiced speech separation. It is based on pitch estimation and comb filter. Indeed, the authors decomposed the composite speech signal into time–frequency units. Then, these *T*–*F* units are used to determine the pitch contour according to theirs zero crossing rate. After that, a comb filter is used to label each unit as target speech or as intrusion.

Finally, we mention Quan et al.'s approach [39] that separates the voice of a target speaker from composite signal by making use of a reference signal from the target signal. They realize this by training two separate neural networks:

- A speaker recognition network that produces speaker-discriminative embeddings;
- A spectrogram-masking network that takes both noisy spectrogram and speaker embedding as input and produces a mask.

We have compared our SNR results from the VCTK database with this method.

Comparing signals (Fig. 12a–c), we conclude that the proposed method is as efficient as those using CASA system.

Figure 12 exposes a composite signal (male and female voices) followed by the reference target signal and its estimated signal using our method and those of Wang–Brown and Hu–Wang. The separated signal is approximately similar to the reference one showing the performance of these methods.

For the entire studied mixture signal formed by male and female voices, the calculation of the signal-to-noise ratio (SNR) gives the following results for the four mentioned approaches.

Table 1 shows that the Zhang–Liu and Hu–Wang systems are the best for the target separation. However, our contribution for the target separation has produced considerable results by using only simple techniques.

In Table 2, we present the computing time of our method, of Hu–Wang system and of Zhang–Liu system.

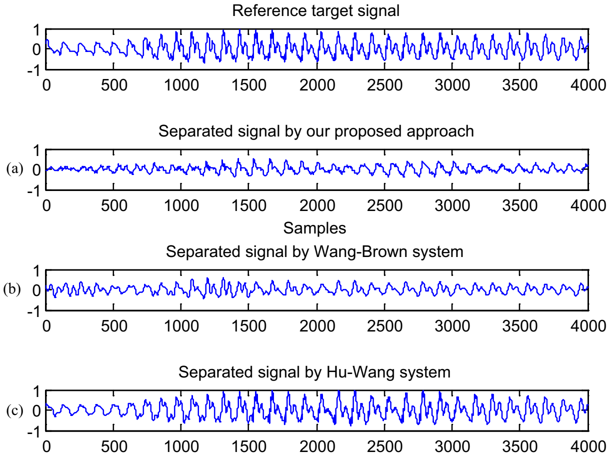

Fig. 12 Reference target signal followed by the estimated target, respectively, by **a** our proposed approach, **b** Wang–Brown system and **c** Hu–Wang system

Table 1 SNR results for speech separation

Approach	Proposed approach	Wang–Brown system [36]	Hu–Wang system [37]	Zhang–Liu system [38]
SNR (dB)	1.97	2.66	10.20	12.91

Table 2 Computing time

	Hu–Wang system [37]	Zhang–Liu System [38]	Our approach
Run time (s)	1064	375	1.72

Table 3 SNR results

	Proposed approach	Quan et al.'s approach [39]
SNR_AVG (dB)	1.03	9.80

From Table 2, we note that our proposed approach has the best computing time. In fact, the computing time of the Hu and Wang model and Zhang–Liu model is much higher compared to our computing time.

For the studied mixture signal formed by male and female voices from the VCTK database, the calculation of the signal-to-noise ratio (SNR) gives the following results (Table 3).

Quan et al.'s system [39] is the best for the target separation since they use a complex and more efficient method based on the training of two separate neural networks. However, our contribution for the target separation has produced significant results by using only simple techniques (MP and filtering).

5 Conclusion

In this work, we presented a composite sound separation algorithm by a designed comb filter tuned by the fundamental frequency f_0. The f_0 is determined by analyzing "the autocorrelation of multi-scale product (AMP)." The composite signal is formed by a mixing of two speech signals, from two male speakers or a male speaker and another female. The "AMP analysis" method estimated peaks for measuring the mean value of the dominant speaker pitch. This approach allows the separation of all voiced sounds in the mixture.

As prospects, we want to ameliorate this approach and to separate a mixture signal composed from more than two speakers. In addition, we would like to separate unvoiced composite signals.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Cao XR, Liu RW (1996) General approach to blind source separation. IEEE Trans Audio Speech Lang Process 44(3):562–571
2. Sakai Y, Mitsuhashi W (2008) A study on the property of blind source separation for preprocessing of an acoustic echo canceller. In: SICE annual conference, pp 13–18
3. Dam H, Cantoni A, Nordholm S, Teo KL (2008) Second order blind signal separation for convolutive mixtures using conjugate gradient. IEEE Signal Process Lett 15:79–82
4. Moreau E, Pesquet JC, Thirion-Moreau N (2007) Convolutive blind signal separation based on asymmetrical contrast functions. IEEE Trans Signal Process 55(1):356–371
5. Won YG, Lee SY (2008) Convolutive blind signal separation by estimating mixing channels in time domain. Electron Lett 44(21):1277–1279
6. Hammam H, Abou Elazmand A, Elhalawanyand M, Abd-El-Samie F (2010) Blind separation of audio signals using trigonometric transforms and wavelet denoising. Int J Speech Technol 13:1–12
7. Boll SF (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Audio Speech Lang Process 27(2):113–120
8. Ephraim Y, Van Trees HL (1995) A signal subspace approach for speech enhancement. IEEE Trans Audio Speech Lang Process 3:251–266
9. Sameti HH, Sheikhzadeh H, Deng L, Brennan RL (1998) HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Trans Audio Speech Lang Process 6:445–455
10. Jensen J, Hansen JHL (2001) Speech enhancement using a constrained iterative sinusoidal model. IEEE Trans Audio Speech Lang Process 9:731–740
11. Brown GJ, Cooke M (1994) Computational auditory scene analysis. Comput Speech Lang 8(4):297–336
12. Wang DL, Brown G (1999) Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans Neural Netw 10(3):684–697
13. Brown GJ, Wang D (2005) Separation of speech by computational auditory scene analysis. In: Benesty J, Makino S, Chen J (eds) Speech enhancement. Springer, Berlin, pp 371–402
14. Wang D, Brown GJ (2006) Computational auditory scene analysis: principles, algorithms, and applications. Wiley IEEE Press, New York
15. Radfar MH, Dansereau RM, Sayadiyan A (2007) Monaural speech segregation based on fusion of source-driven with modeldriven techniques. Speech Commun 49(6):464–476
16. Bregman AS (1994) Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge
17. Klapuri A (2003) Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. IEEE Trans Speech Audio Process 11(6):204–816
18. Walmsley PJ, Godsill SJ, Rayner PJW (1999) Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In: Proceedings of IEEE workshop on audio and acoustics, Mohonk, NY State Mohonk, NY State
19. Tolonen T, Karjalainen M (2000) A computationally efficient multipitch analysis model. IEEE Trans Speech Audio Process 8(6):708–716
20. Godsill S, Davy M (2002) Bayesian harmonic models for musical pitch estimation and analysis. In: Proceedings of ICASSP, pp 1769–1772
21. Wu M, Wang D, Brown GJ (2003) A multi-pitch tracking algorithm for noisy speech. In: ICASSP, pp 369–372
22. Miwa T, Tadakoro Y (2000) The problems of transcription using comb filters for musical instrument sounds and their solutions. Technical Report of IEICE
23. Gainza M, Lawlor B, Coyle E (2004) Harmonic sound source separation using FIR comb filters. In: 117th AES convention, San Francisco
24. Zeremdini J, Ben Messaoud MA, Bouzid A (2017) Multiple-comb filters and autocorrelation of the multi-scale product for multi-pitch estimation. Appl Acoust 120:45–53
25. Zeremdini J, Ben Messaoud MA, Bouzid A (2017) Evaluation of multipitch estimation by multi-scale product analysis. In: Conférence internationale IEEE ICEMIS 2017, Monastir, Tunisie, 8–10 Mai 2017
26. Zeremdini J, Ben Messaoud MA, Bouzid A (2017) Multi-pitch estimation based on multi-scale product analysis, improved comb filter and dynamic programming. Int J Speech Technol 20(2):225–237
27. Zeremdini J, Ben Messaoud MA, Bouzid A (2015) A comparison of several computational auditory scene analysis (CASA) techniques for monaural speech segregation. Brain Inform 2:155–166
28. Zeremdini J, Ben Messaoud MA, Bouzid A, Ellouze N (2013) Contribution to the multi-pitch estimation by multi-scale product analysis. NOLISP, Mons, Belgique
29. Sambur MR (1978) Adaptive noise canceling for speech signals. IEEE Trans Acoust Speech Signal Process 26:5
30. Darlington DJ, Campbell DR (1996) Sub-band adaptive jittering applied to speech enhancement. In: ESCA tutorial and research workshop, the auditory basis of speech perception, Kccle University, pp 15–19

31. Iliev G, Kasabov N (1998) Adaptive filtering with averaging in noise cancellation for voice and speech recognition. University of Otago, Dunedin
32. Christensen MG, Jakobsson A (2010) Optimal filter designs for separating and enhancing periodic signals. IEEE Trans Signal Process 58:12
33. Cooke M, Barker J (2006) An audio-visual corpus for speech perception and automatic speech recognition. J Acoust Soc Am 120(25):2421–2424
34. Veaux C, Yamagishi J, MacDonald K et al (2017) Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://datashare.is.ed.ac.uk/handle/10283/2651
35. https://google.github.io/speaker-id/publications/VoiceFilter/
36. Wang DL, Brown GJ (1999) Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans Audio Speech Lang Process 10:684–697
37. Hu G, Wang DL (2010) A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans Audio Speech Lang Process 18(8):2067–2079
38. Zhang X, Liu W (2011) Monaural voiced speech segregation based on pitch and comb filter. In: Conférence internationale Interspeech 2011, Florence, Italy, 28–31 August 2011
39. Wang Q, Muckenhirn H, Wilson K, Sridhar P, Wu Z, Hershey J, Saurous RA, Weiss RJ, Jia Y, Moreno IL (2019) Voicefilter: targeted voice separation by speaker-conditioned spectrogram masking. Cornell University, Ithaca

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.