A Survey on Single Channel Speech Separation

G. Logeshwari and G.S. Anandha Mala

Department of Computer Science and Engineering,
St. Joseph's College of Engineering,
Chennai – 600 119, Tamil Nadu, India
logesh_gd@yahoo.com,
gs.anandhamala@gmail.com

Abstract. Single channel speech separation is a branch of speech separation process, which is an ongoing interesting research topic for the past 40 years and continues till now, but still there is a lack in separating the required signal from the mixture of signals with 100% accuracy and be used by the common people. Many researches have been done in various ways using the parameters like pitch, phase, magnitude, amplitude, frequency and energy, spectrogram of the speech signal. Various issues in single channel speech separation process are surveyed in this paper and the major challenges faced by the speech research community in realizing the system are pointed out as conclusion.

Keywords: Computational Auditory Scene Analysis, Independent component Analysis, Amplitude, Pitch.

1 Introduction

In our daily life we hear sounds not in isolation but in mixture with background noise which depends on the environment like car noise, television noise, radio noise and crowd noise called as cocktail party effect. We humans have the capability to recognize the target speech eliminating the back ground noise. But as a system it will capture the combinations of several speech signals as a mixture which overlaps in time and frequency. Single channel speech separation means separation of a specific, required speech signal from a mixture of speech signals or from background noise, where the speech mixture is captured by a single microphone. Single channel speech separation is also called as Multiple Input Single Output System (MISO) which is a branch of Speech Separation Process.

The pioneering work of separating audio signals from the back ground noise starts from 1976 and the research continues till now. The main goal in speech separation research is to model the processes underlying the mechanism of separation by humans and replicate it in machines. Various techniques have been proposed for single channel source separation, using either 'time domain' or 'frequency domain' technique. When the noisy speeches are recorded in strong reverberant environment, the time domain method needs to learn many parameters and may result in convergence difficulty and heavy computation load. The frequency domain method,

V.V. Das and J. Stephen (Eds.): CNC 2012, LNICST 108, pp. 387–393, 2012.

in contrast, can reduce the convolutive mixtures to complex valued instantaneous mixtures in each frequency bin, and thus has much simplified computation and faster convergence compared to the time domain one. Therefore this paper mainly concentrates on frequency domain speech separation algorithms when compared to time domain speech separation algorithms.

This paper is organized as follows. Section 2 deals the broad classification of single channel speech separation systems, under which the research has been carried out and various techniques employed for processing single channel speech corrupted by additive and / or convolutive noise. Section 3 concludes the paper by suggesting future directions for further development for the growth of this field.

2 Classification of Speech Separation

Speech separation process is broadly classified as i) Equal Input Equal Output System (EIEO), ii) Multiple Input Single Output System (MISO also called as Single channel speech separation – SCSS), and iii) Multiple Input Multiple Output System (MISO). This paper concentrates on MISO research issues. The two main approaches of Single channel speech separation are Independent component Analysis (ICA) approach using signal processing techniques and Computational Auditory Scene Analysis (CASA) approach using auditory modeling techniques. Both the techniques can be solved in time domain and frequency domain. The frequency domain has been further subdivided into pitch based algorithms and amplitude based algorithms.

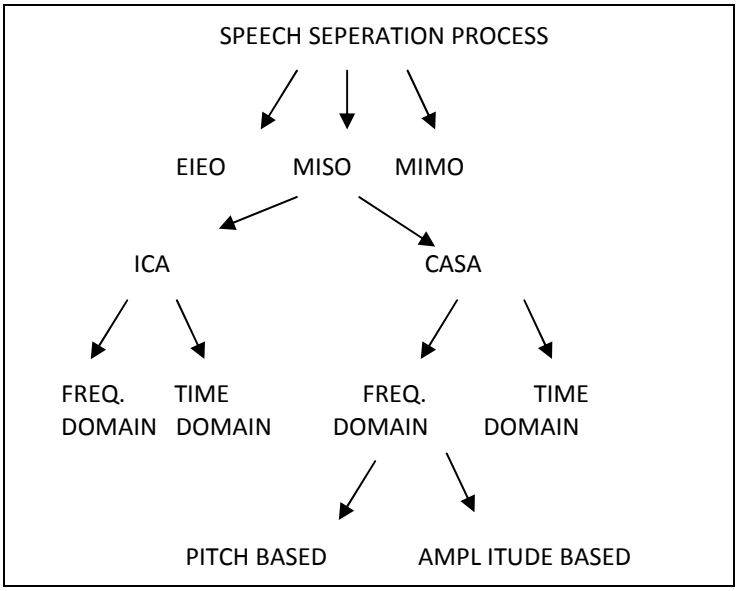

Fig. 1. Overview of Speech Separation Process

2.1 ICA Based Speech Separation System

Independent component analysis (ICA) consisted of recovering a set of maximally independent sources from the observed mixtures without the knowledge of the source signals and the mixing parameters and thus provided effective solutions for blind source separation (BSS). Since ICA required little prior information about the source signals and mixing parameters, it had become a widely used method for blind separation of noisy speech signals. Thomas W.Parsons [1] concentrated on separation of vocalic speech and the segregation of concurrent voices by frequency domain approach, recorded by single channel.

2.2 CASA Based Speech Separation System

The monaural system segregating voices from two talkers proposed by Weintraub [2] might be considered as the first Computational Auditory Scene Analysis (CASA) study. Many approaches based on CASA techniques, employed human auditory model at the front end of a speech separation system and these systems were evaluated by automatic speech recognizers. The performance of such systems was invariably compared with that of human through perception tests. The treatment on CASA approach to speech separation and associated problems were discussed in detail in [3].

Frequency Domain Based Speech Separation System. Most algorithms that dealt with this problem were based on masking, wherein unreliable frequency components from the mixed signal spectrogram were suppressed, and the reliable components were inverted to obtain the speech signal from speaker of interest. Most techniques estimated the mask in a binary fashion, resulting in a hard mask. In [4], Aarthi estimated all the spectral components of the desired speaker and estimated a soft mask that weights the frequency sub bands of the mixed signal. This algorithm was computationally expensive but achieved better performance than that obtained with hard binary masks. This computational complexity was reduced by deriving a soft mask filter, using minimum square error estimation of the log spectral vectors of sources which were modeled using the Gaussian composite source modeling approach[5].

The separation using instantaneous frequency from each narrow band frequency channel using short-time Fourier analysis based on cross correlation was better than the separation based on fundamental frequency alone [6]. Yun used both the magnitude and phase information of the signal to extract and enhance the desired speech signal from the mixed speech signals [7]. The complexity of STFT was reduced by using sinusoidal parameters composed of amplitude and frequency [8, 9] and this algorithm worked independently for pitch estimation.

An adaptive time-frequency resolution approach for non negative matrix factorization was proposed by Serap[10] to improve the quality and intelligibility of the separated sources. The degradation due to non negative matrix factorization on a complex valued short time Fourier transform was reduced by incorporating phase estimates via complex matrix factorization[11]. The speech mixture was

decomposed into a series of oscillatory components and derived a sparse non negative matrix factorization to estimate the spectral bases and temporal codes of the sources. This methodology required no training knowledge for speech separation[12]. Least Squares fitting approach method was proposed by Srikanth to model the speech mixture as a sum of complex exponentials in which, it separated the participating speaker streams rather than in the favor of the dominating speaker [13].

Amplitude Modulation Based Algorithm. The time domain signal was first transformed into a time-frequency representation by applying Short Time Fourier Transform. Then, the instantaneous amplitude was calculated for each narrow band frequency bin. As the mixed speech signal had a great deal of overlap in the time domain, modulation frequency analysis could provide a greater degree of separation among sources [14]. A segment of the speech mixture was sparsely decomposed into periodic signals, with time varying amplitude in which each of them was a component of the individual speaker. For speech separation, the author [15] used K means clustering algorithm for the set of the periodic signals. After the clustering, each cluster was assigned to its corresponding speaker using codebooks that contained spectral features of the speakers which could perform with less computational cost.

Pitch Tracking Based Algorithm. It was very natural to imagine that speech separation could be accomplished by detecting the pitch of the mixed speech. Generally speaking, pitch estimation could be done using either temporal, spectral or spectro temporal methods (e.g., [16], [17], [18]). To identify the pitch contours of each of several simultaneous speakers, comb filtering or other techniques could be used to select the frequency components of the target speaker and suppress other components from competing speakers.

The autocorrelation function of cochlear outputs was computed using dynamic programming to estimate the dominant pitch. The components of dominating speaker were removed and this process was repeated to retrieve the pitch values for the weaker speaker [19]. Though, simple and easy to implement, it did not lead to a satisfactory reduction in word error rate. Other researchers ([20], [21]) had proposed similar recursive cancelation algorithms in which the dominant pitch value was first estimated, and then removed so that a second pitch value could be calculated. All of these algorithms were critically dependent on the performance of the first estimation stage, and errors in the first pass usually led to errors in all subsequent passes. The signal of target speaker was separated from an interfering speaker by manually masking out modulation spectral features of the interferer. But this algorithm needed a rough estimate of the target speaker's pitch range [22]. Hence Mahmoodzadeh estimated the pitch range in each frame of modulation spectrum of speech by analyzing onsets and offsets. He filtered the mixture signal with a mask extracted from the modulation spectrogram of mixture signal [23].

Hu estimated the multiple pitches present in the mixture simultaneously from the speech signal and performed voiced/unvoiced decisions at the same time by separating speech into low and high frequency segments [24]. For multi pitch estimation, Michael Stark utilized the factorial HMM method. He modeled the vocal tract filters either by vector quantization or by non negative matrix factorization for

the fast approximation for the likelihood computation [25]. The improvement on speech quality was consistent with Ji's conclusion that long speech segments maintained the temporal dynamics and speaker characteristics better than short segments[26] In a long-short frame associated harmonic model, the long frame could achieve high harmonic resolution, while the short frame could ensure the short time stationary feature of the speech signal. They were jointly used to improve the accuracy of the multi pitch estimation [27].

Time Domain Based Speech Separation System. Majority of research articles reviewed so far used 'frequency domain' for developing any algorithm for speech separation. As the spectral techniques assumed that source signals were disjoint in the spectrogram, few implementations resulted in the audible distortions of the signal. So a refiltering technique was estimated using time varying mask filters that localized sound streams in the spatio temporal region was proposed in [28][29]. High level of separation performance had been demonstrated from simulated single channel recordings, by exploiting a priori sets of time domain basis functions learned by ICA to the separation of mixed source signals [30] and this method was demonstrated for real world problems such as blind source separation, denoising. Without modeling individual speakers, an adaptive, speech specific segmentation algorithm using spectral learning approach had been used to separate the speech mixtures [31]. By exploiting the inherent time structure of sound sources, Gil learned a priori sets of time domain basis functions that encode the sources in a statistically efficient manner using a maximum likelihood approach [32]. The technique of time-domain decomposition of signal into pulses of amplitude modulation, demonstrated that, signals low-pass filtered in the modulation domain maintained bursts of energy which were comparable to those that could be extracted entirely within the time-domain. Garreth focused on the instantaneous features of transient changes in loudness [33].

3 Conclusion

The paper highlights the importance of single channel speech separation systems critically reviewed its growth in the last few decades, raising an awareness of the challenges faced by the researchers in the development of new theory and algorithms. In separating the speech signals, a priori knowledge of the underlying sources are used to estimate the sources. Hence a system should be designed for separating two sources without the prior knowledge of the source signals. Also the overall performance of the system degrades if it is for speaker independent source separation system. There is a lack in separating a more accurate speaker separation system without any cross talk which is mainly due to the interference of unvoiced segments in the target signal. Hence algorithms are to be designed to deal with unvoiced segments. The long short associate harmonic model can handle voiced and unvoiced speech separation when it is mixed with the voiced speech by detecting the energy of the high frequency. However, if two unvoiced speech signals occur simultaneously, it fails. Hence the proposed system is to combine the long short frame associated method with the clustering algorithm to handle the inharmonic structures of unvoiced

segments. This paper summarizes by suggesting new directions of research to be focused in future by researchers concurrently working in their respective fields in order to eventually see single channel speech separation based end products that will be extremely useful to the community at large.

References

1. Parson, T.W.: Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Am. 60(4), 911–918 (1976)
2. Weintraub, M.: A theory and computational model of Auditory Monaural Sound Separation. Ph.D Thesis, Stanford University (1985)
3. Wang, D.L., Brown, G.J.: Computational Auditory Scene Analysis. John Wiley&Sons (2006)
4. Reddy, A.M., Raj, B.: Soft Mask Methods for Single Channel Speaker Separation. IEEE Tran. Audio, Speech, Lang. Process. 15(6), 1766–1776 (2007)
5. Radfar, M.H., Dansereau, R.M.: Single Channel Speech Separation Using Soft Mask Filtering. IEEE Tran. Audio, Speech, Lang. Process. 15(8), 2299–2310 (2007)
6. Gu, L.: Single-Channel Speech Separation based on Instantaneous Frequency, Carnegie Mellon University, Ph.D Thesis (2010)
7. Lee, Y.-K., Lee, I.S., Kwon, O.-W.: Single Channel Speech Separation Using Phase Based Methods. Procedures of the IEEE Tran. Acoust., Speech, Signal, Process. 56(4), 2453–2459 (2010)
8. Mowlaee, P., Christensen, M.G., Jensen, S.H.: New Results on Single-Channel Speech Separation Using Sinusoidal Modeling. IEEE Tran. Audio, Speech, Lang. Process. 19(5), 1265–1277 (2011)
9. Mowlaee, P., Saeidi, R., Tan, Z.H., Christensen, M.G., Kinnunen, T.: Sinusoidal Approach for the Single Channel Speech Separation and Recognition Challenge. In: Proc. Interspeech, pp. 677–680 (2011)
10. Kırbız, S., Smaragdis, P.: An adaptive time-frequency resolution approach for non-negative matrix factorization based single channel sound source separation. In: Proc. IEEE Conference ICASSP, pp. 253–256 (2011)
11. King, B.J., Atlas, L.: Single-Channel Source Separation Using Complex Matrix Factorization. IEEE Tran. Audio, Speech, Lang. Process. 19(8), 2591–2597 (2011)
12. Gao, B., Woo, W.L., Dlay, S.S.: Single-Channel Source Separation Using EMD-Subband Variable Regularized Sparse Features. Tran. Audio, Speech, Lang. Process. 19(4), 961–976 (2011)
13. Vishnubhotla, S., Espy-Wilson, C.Y.: An Algorithm For Speech Segregation of Co-Channel Speech. In: Proc. IEEE Conference ICASSP, pp. 109–112 (2009)
14. Schimmel, S.M., Atlas, L.E., Nie, K.: Feasibility of single channel speaker separation based on modulation frequency analysis. In: Proc. IEEE Conference ICASSP, pp. IV605–IV608 (2007)
15. Nakashizuka, M., Okumura, H., Iiguni, Y.: Single Channel Speech Separation Using A Sparse Periodic Decomposition. In: Proc. 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, pp. 218–222 (2009)
16. Bach, F., Jordan, M.: Discriminative training of hidden markov models for multiple pitch tracking. In: Proc. of ICASSP, pp. v489–v492 (2005)
17. Charpentier, F.J.: Pitch detection using the short-term phase spectrum. In: Proc. of ICASSP, pp. 113–116 (1986)

18. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals. Prentice-Hall, Englewood (1993)
19. Weintraub, M.: A computational model for separating two simultaneous talkers. In: Proc. of ICASSP, pp. 81–84 (1986)
20. de Cheveigne, A., Kawahara, H.: Multiple period estimation and pitch perception model. Speech Communication 27(3-4), 175–185 (1999)
21. Barker, J., Coy, A., Ma, N., Cooke, M.: Recent advances in speech fragment decoding techniques. In: Proc. of Interspeech, pp. 85–88 (2006)
22. Schimmel, S.M., Atlas, L.E., Nie, K.: Feasibility of Single Channel Speaker Separation Based on Modulation Frequency Analysis. In: Proc. of ICASSP, pp. IV605–IV608 (2007)
23. Mahmoodzadeh, Abutalebi, H.R., Soltanian-Zadeh, H., Sheikhzadeh, H.: Single Channel Speech Separation with a Frame-based Pitch Range Estimation Method in Modulation Frequency. In: Proc. of IST, pp. 609–613 (2010)
24. Hu, G., Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Tran. on Neural Networks 15(5), 1135–1150 (2004)
25. Stark, M., Wohlmayr, M., Pernkopf, F.: Source–Filter-Based Single-Channel Speech Separation Using Pitch Information. IEEE Trans. on Acoustics, Speech, Signal Process. 19(2), 242–255 (2011)
26. Ji, M., Srinivasan, R., Crookes, D.: A corpus-based approach to speech enhancement from nonstationary noise. In: Proc. of Interspeech, Makuhari, Chiba, Japan, pp. 1097–1100 (2010)
27. Huang, Q., Wang, D.: Single-channel speech separation based on long-short frame associated harmonic model. Digital Signal Processing 21, 497–507 (2011)
28. Roweis, S.T.: One microphone source separation. In: Proc. of NIPS-13, pp. 793–799 (2001)
29. Roweis, S.T.: Factorial models and refiltering for speech separation and denoising. In: Proc. Eurospeech, pp. 1009–1012 (2003)
30. Jang, G.J., Lee, T.W.: A maximum likelihood approach to single channel source separation. Journal of Machine Learning Research 4(7-8), 1365–1392 (2004)
31. Bach, F., Jordan, M.I.: Blind one-microphone speech separation: A spectral learning approach. Neural Info. Process. System, 65–72 (2005)
32. Jang, G.-J., Lee, T.-W., Oh, Y.-H.: Single channel Signal Separation Using Time-Domain Basis Functions. IEEE Signal Processing Letters 10(6), 168–171 (2003)
33. Prendergast, G., Johnson, S.R., Green, G.G.R.: Extracting amplitude modulations from speech in the time domain. Speech Communication 53, 903–913 (2011)