

Hybrid Approach to Single-Channel Speech Separation Based on Coherent–Incoherent Modulation Filtering

Azar Mahmoodzadeh¹ · Hamid Reza Abutalebi¹ 

Received: 17 June 2015 / Revised: 10 August 2016 / Accepted: 11 August 2016 /
Published online: 26 August 2016
© Springer Science+Business Media New York 2016

Abstract Single-channel speech separation is a challenging problem that has been of particular interest in recent years. Here the goal is to separate the target speech signal from the interference signals, with high accuracy. We propose a new hybrid single-channel speech separation system that applies adaptive coherent modulation filtering for low-frequency subbands and iterative incoherent speech separation technique for high-frequency subbands. In the adaptive coherent modulation filtering, an affine projection filter is applied to subband envelope in order to eliminate the interference signal. The subband envelope is determined via demodulation of the subband signal using a coherently detected subband carrier based on the time-dependent spectral center-of-gravity demodulation. The adaptive affine projection filter uses the separated target signal obtained from the iterative incoherent speech separation system as a reference signal. This system first obtains a rough estimate of target fundamental frequency range and then uses this estimate to segregate target speech. It then improves both fundamental frequency range estimation and voiced speech separation iteratively. Perceptual evaluation of speech quality, as one of the evaluation indices investigated in this paper, indicates that the proposed system extracts the majority of target speech segments with minimal interference and outperforms previous systems in voiced speech separation.

✉ Hamid Reza Abutalebi
habutalebi@yazd.ac.ir

Azar Mahmoodzadeh
a.mahmoodzadeh@stu.yazd.ac.ir

¹ Department of Electrical Engineering, Yazd University, Pajuhesh St., Safaieh,
Postal Box: 89195-741, Yazd, Iran

Keywords Single-channel speech separation · Coherent and incoherent demodulation · Modulation filtering · Carrier estimator · Instantaneous frequency · Modulator signal

1 Introduction

In real-world environments, speech reaching our ears is almost never pure. Corruption of speech by various types of acoustic interference, such as fan noise, music, or other voices, poses a serious problem for many applications including automatic speech recognition, telecommunication systems, and hearing aid design. While separating a particular sound from a mixture of many sources is a major challenge for machines, the human auditory system is able to perform such a task well, even in difficult and cluttered acoustical environments. Since speech may be corrupted by various types of noise and interferences, single-channel speech separation (SCSS) is one of the most difficult problems among the speech enhancement methods. Separating speech from interference using a single microphone is a particularly difficult task because only one recording is available and one cannot explore the spatial information of sources present in multi-microphone situations.

SCSS methods can be categorized into the following groups: speech modeling, non-negative matrix factorization (NMF), computational auditory scene analysis (CASA), and modulation filtering [30].

Model-based methods separate speech from interference by modeling source patterns and formulating the separation as an estimation problem in a probabilistic framework. Such methods estimate individual speech utterances or derive a time-frequency (T-F) mask for segregation using source models [10, 17, 29]. To train the models in these methods, vector quantization (VQ), Gaussian mixture models (GMM), and hidden Markov models (HMM) are used. These methods work based on the prior knowledge; accordingly, the performance of these systems is reduced when they are applied to a new noise and speech signal condition. Therefore, they are inappropriate in real-life applications.

In the NMF method, a nonnegative matrix, which is the representation of the noisy speech signal, is factorized into two nonnegative matrices [6, 24]. Unfortunately, for highly overlapped sources, the performance of this method is not so good.

CASA has borrowed its considerable organization from perceptual principles; hence, it does not use any a priori knowledge of the speakers in speech separation [32, 33]. CASA procedure can be implemented in two main stages: segmentation and grouping [25, 39, 41]. In the segmentation stage, the input is decomposed into several contiguous T-F segments originating mainly from a single sound source. The grouping stage sequentially combines segments that are likely to be generated from the same source. Fundamental frequency is an important factor for speech separation in CASA methods. But in the presence of interfering speech signals, the performance of multi-pitch trackers degrades; this, in turn, deteriorates the performance of CASA method [22, 43].

Modulation analysis and filtering have been useful for modifying narrowband analytic signals and source separation [2, 12, 15, 23]. Atlas et al. [4] and Schimmel et al.

[38] have used modulation filtering to separate sound mixtures of musical instruments based on different characteristics of the instrument sounds in modulation frequency domain. In these methods, prior knowledge of the interference signal is required to design a modulation filter; however, in practice there is no access to interference signal in single-channel recording [16]. In [37], by assuming that the fundamental frequency range is known and this range is constant in each frequency band, modulation spectral analysis is used as a tool for producing a speech separation mask in modulation spectrogram domain.

In modulation filtering methods, an input signal is divided into subband signals via a filter-bank or a short-time Fourier transform (STFT). Then, each subband is decomposed into a slowly varying modulator and a narrow-band carrier signal. Each subband modulator signal is passed through a linear time-invariant filter and subsequently multiplied by the original unmodified subband carrier signal to obtain the modulation filtered subband signals. Finally, a broadband modulation filtered output signal is reconstructed by summing the filtered subbands.

Decomposition of a subband signal into its modulator and carrier using the Hilbert transform [12] or a direct magnitude estimate of the modulator [40] is called incoherent method [5,36]. It has been shown that the incoherent method causes distortion and reduces the effectiveness of modulation filters [14,36]. This distortion occurs because of three important limitations: (1) subband magnitude and phase signals usually exceed bandwidth of the subband signal; (2) incoherent detectors force a conjugate symmetric spectrum on the modulator, which is an unrealistic assumption for most natural signals; (3) modulator domain of incoherent detectors is not closed under convolution [35]. This distortion can be avoided to a large degree by estimating a modulator and a carrier using a coherent method (see [4,36,38] for details). The coherent approach decomposes a subband signal into a modulator and a carrier by estimating a subband carrier signal using an instantaneous frequency (IF) estimator and coherently demodulating the subband signal by the carrier to obtain the subband modulator.

Motivated by the limitations of the methods summarized above, Mahmoodzadeh et al. [28] proposed a combination of coherent and incoherent methods to develop adaptive modulation filtering for voiced speech separation in a single-channel scenario. In the system proposed in that paper, the subband signal is decomposed into the modulator and carrier signals using an estimate of the IF based on the time-dependent spectral center-of-gravity (COG), introduced by Clark [8]. In modulation filtering, an analytic subband signal is modified by filtering the modulator and combining the result with the original carrier. We apply an adaptive affine projection (AP) filter on each subband modulator signal to separate the target speech from the interference signal. This idea was presented in a preliminary form in [28], and now it is extended and modified in this paper.

For the adaptive coherent modulation filtering, the reference channel is needed but this is not available in single-channel recordings. To solve this problem, we propose an iterative incoherent speech separation system to approximate the target speech signal. The proposed method estimates the target fundamental frequency range and segregates the target in tandem. It obtains a rough estimate of the target fundamental frequency range and uses this estimate to segregate the target speech. With the segregated target, we generate an accurate estimate of the fundamental frequency range and use it to

improve the separation process. This means that the proposed system performs the fundamental frequency range estimation and speech separation jointly and iteratively.

It is known that high-frequency critical bands have wider bandwidths compared to low-frequency bands. Therefore, higher harmonics of speech signals have wider bandwidth than lower harmonics [35]. This leads to difficulties for coherent carrier estimators in correctly estimating carriers in high-frequency subbands. To avoid this problem, in our proposed hybrid system, we employ coherent modulation filtering for low-frequency subbands and incoherent modulation filtering for high-frequency subbands.

We assess performance of the proposed method for different intrusion types and signal-to-noise ratio (SNR) levels. The results show that the proposed method performs well in single-channel speech separation.

The rest of the paper is organized as follows. Section 2 first describes coherent and incoherent demodulation methods and then presents the proposed hybrid system. Systematic evaluation and comparison studies are described in Sect. 3, and conclusions are made in Sect. 4.

2 Proposed Speech Separation System

The block diagram of the proposed multi-stage system is shown in Fig. 1. In the first stage, the wideband noisy speech signal is decomposed into a set of narrow-band signals employing a uniform filterbank. Subsequently, narrowband noisy speech signal is divided into modulator and carrier signals using the coherent demodulation method. Then, the target signal is separated from interference using adaptive coherent modulation filtering. In this stage, an adaptive AP filter is applied on each subband modulator.

This filter requires a reference channel which is not available in single-channel recordings. To solve this problem, as shown in Fig. 1, we use an estimated target signal obtained from our proposed iterative incoherent speech separation system described below. As shown in the last stage of Fig. 1, the hybrid system is used to separate

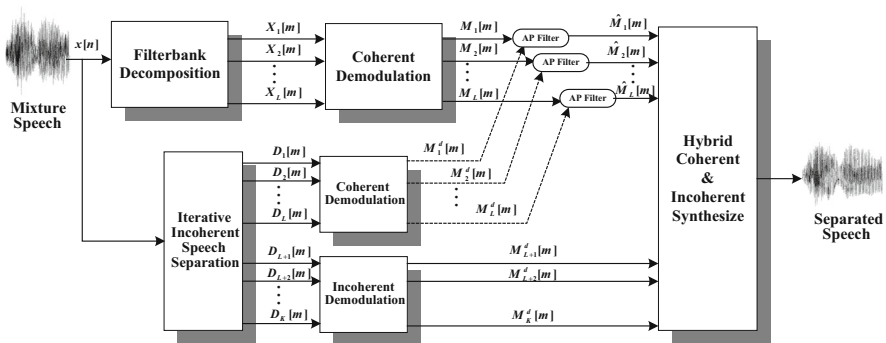


Fig. 1 Block diagram of the proposed hybrid coherent–incoherent system for single-channel speech separation

the high- and low-frequency subbands. A detailed description of each stage and the corresponding signals are presented in the following subsections.

2.1 Coherent and Incoherent Demodulation

The first step in modulation analysis and filtering is the decomposition of a narrow-band signal into carrier and modulator (or envelope) signals [35]. Modulator is a low-pass signal describing amplitude modulation of the signal, while carrier signal represents frequency modulation of the narrowband signal. We denote a wideband discrete-time signal by $x[n]$ where n represents the discrete-time index. STFT generates a T-F representation of the signal $x[n]$. The narrowband signal $X_k[m]$ is written as:

$$X_k[m] = \sum_{n=0}^{K-1} x[n] w[mM - n] e^{-j2\pi nk/K} \quad k = 0, \dots, K - 1 \quad (1)$$

where k denotes the subband (channel) index, m represents the time index, M is the decimation factor, K represents the STFT length (equal to the number of subbands in the filter bank), and $w[\cdot]$ denotes the analysis window with length L . STFT provides a uniform filterbank for decomposing a broadband signal into a set of narrowband subband signals.

Demodulation is defined as the process of estimating $M_k[m]$ and $C_k[m]$ for a given signal $X_k[m]$ for all m and k . The signal product model of the k^{th} analytic band-pass signal $X_k[m]$ can be expressed as:

$$X_k[m] = M_k[m] \cdot C_k[m] \quad (2)$$

where $M_k[m]$ and $C_k[m]$ are the modulator and carrier signals of the k^{th} subband, respectively; these signals are obtained using either incoherent or coherent demodulation approaches as follows.

In the incoherent demodulation method [35], the signal's modulator is found by applying an envelope detector to the signal as:

$$M_k[m] \triangleq \mathfrak{D} \{X_k[m]\} \quad (3)$$

where \mathfrak{D} is the envelope detector operator. For complex-valued subbands $X_k[m]$, the incoherent modulator signal is actually obtained using the magnitude operator [12]:

$$M_k[m] = |X_k[m]| \quad (4)$$

Having $M_k[m]$, the carrier signal is calculated from (2).

On the other side, in the coherent demodulation, the signal's carrier is achieved by applying a carrier detector to the signal as:

$$C_k[m] \triangleq \mathfrak{D}^c \{X_k[m]\} \quad (5)$$

where \mathfrak{D}^c is the carrier detector operator. Based on the COG method [3, 8], in coherent demodulation method, the complex carrier is defined as:

$$C_k[m] = \exp(j\phi_k[m]) \quad (6)$$

where $\phi_k[m]$ is the phase of the carrier $C_k[m]$. Thus, the main step in computing the coherent modulator is to find the phase of the carrier, $\phi_k[m]$. The IF of the oscillating signal $C_k[m]$ is defined as the derivative of the phase. Therefore, the phase is calculated via the cumulative sum of the carrier frequency $f_k[m]$ as:

$$\phi_k[m] = \sum_{q=0}^m f_k[q] \quad (7)$$

Hence, determination of the IF of the subband is essential in computing the modulator and carrier signals. In the COG approach, $f_k[m]$ is defined as the average frequency of the instantaneous spectrum of $X_k[m]$ at time m as:

$$f_k[m] = \frac{\sum_{r=0}^{R-1} z[r] |X_k[r, m]|^2}{\sum_{r=0}^{R-1} |X_k[r, m]|^2}, \quad (8)$$

where $z[r]$ is a weighting function defined as:

$$z[r] = \begin{cases} 2\pi r/R, & 0 \leq r \leq R/2 \\ 2\pi r/R - 2\pi, & R/2 < r < R \end{cases}, \quad (9)$$

and the instantaneous spectrum, $X_k[r, m]$, is estimated as:

$$X_k[r, m] = \sum_{p=0}^{R-1} g[p] X_k[m+p] e^{-j2\pi(r/R)p} \quad r = 0, \dots, R-1. \quad (10)$$

where $g[n]$ denotes the analysis window with length R .

2.2 Adaptive Coherent Modulation Filtering

The main idea of this stage of the proposed system is the separation of target speech signal from the interference based on adaptive coherent modulation filtering. As depicted in the multi-stage block diagram of Fig. 1, the first step in the adaptive coherent modulation filtering is determination of the modulator and carrier parts of the noisy subband signal using the COG demodulation method. Coherent detectors are able to adjust the bandwidth of the carrier by estimating the IF of the signal and thereby adjust the bandwidth of the modulator. Therefore, the complex modulator signal is closed under convolution [35].

To separate the target from the interference signal, each subband modulator signal of noisy speech is filtered using the adaptive AP filter as shown in Fig. 1. In each subband, the filter coefficients of the adaptive filter are determined by the subband affine projection algorithm. The LMS algorithm and its variants, that are traditionally

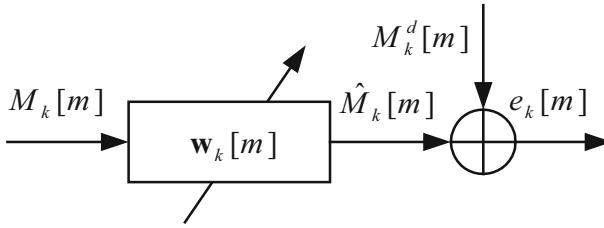


Fig. 2 Prototype adaptive filter setup

the most common adaptation algorithms, are problematic in the sense of convergence speed, especially in the case of colored inputs. Instead, AP converges to the true value monotonically and much faster compared to the LMS [1, 34]. Due to these reasons, AP is chosen in this paper.

Figure 2 shows a prototype adaptive filter setup where $\mathbf{M}_k[m]$, $\mathbf{M}_k^d[m]$, and $\mathbf{e}_k[m]$ are the input, the desired, and the output error signal vectors, respectively. The vector $\mathbf{w}_k[m]$ is the $M \times 1$ column vector of filter coefficient at time m , in such a way that the output, $\hat{M}_k[m]$, is an estimate of the desired signal, $M_k^d[m]$. The filter vector update equation for the AP algorithm is given by:

$$\mathbf{w}_k[m + 1] = \mathbf{w}_k[m] + \mu \mathbf{Y}_k[m] \mathbf{C}_k^{-1}[m] \mathbf{e}_k[m], \tag{11}$$

where μ is the step size that determines the convergence speed, and the input matrix $\mathbf{Y}_k[m]$ is given by:

$$\mathbf{Y}_k[m] = [\mathbf{y}_k[m], \mathbf{y}_k[m - 1], \dots, \mathbf{y}_k[m - p + 1]], \tag{12}$$

and p is the order of AP algorithm. The columns of matrix $\mathbf{Y}_k[m]$ are furthermore defined through

$$\mathbf{y}_k[m] = [M_k[m], M_k[m - 1], \dots, M_k[m - M + 1]]. \tag{13}$$

Also, the output error signal, $\mathbf{e}_k[m]$, and the autocorrelation matrix, $\mathbf{C}_k[m]$, are given by:

$$\mathbf{e}_k[m] = \mathbf{M}_k^d[m] - \mathbf{Y}_k[m] \mathbf{w}_k[m], \tag{14}$$

and

$$\mathbf{C}_k[m] = \mathbf{Y}_k^H[m] \mathbf{Y}_k[m] + \delta \mathbf{I}, \tag{15}$$

where the scalar δ is a regularization parameter and the vector of desired signal is given by:

$$\mathbf{M}_k^d[m] = [M_k^d[m], M_k^d[m - 1], \dots, M_k^d[m - p + 1]]. \tag{16}$$

Theoretically, the desired modulator signal, $M_k^d[m]$, should be obtained by applying the analysis filterbank on the desired signal $d[n]$ and then decomposing the subband

signal $D_k[m]$ into the carrier and modulator signals $C_k^d[m]$ and $M_k^d[m]$. Obviously, a replica or an estimate of the desired signal is necessary for adaptive coherent modulation filtering. Since the desired signal is not directly available in the single-channel case, as shown in Fig. 1, we employ the proposed iterative incoherent speech separation system to obtain an estimation of the target speech signal.

2.3 Iterative Incoherent Speech Separation System

The main objective of iterative incoherent speech separation system is to separate the target signal from the interference signal and estimate a relatively clean signal as a reference channel for adaptive coherent modulation filtering. This single-channel speech separation system is implemented using a soft mask in the modulation spectrogram domain. To make the soft mask, the determination of the fundamental frequency range of the target speaker is necessary.

Precise estimation of the fundamental frequency range has detrimental effect on the performance of the soft mask in separating the target signal from the interfering ones. Nevertheless, interference signals reduce the performance of the fundamental frequency range estimation algorithm. To overcome this limitation, we propose an iterative system for speech separation.

In the proposed system, we first determine a rough estimate of the fundamental frequency range in the modulation frequency domain based on the modulation spectrogram of the noisy speech. Speech separation is performed using the soft mask obtained from the modulation spectrogram based on this initial estimate of the fundamental frequency range. The value of this mask in each subband depends on the estimated fundamental frequency range in the same subband.

The separated signal is fed back to the input of the iterative system as a noisy signal. Since the SNR of the new noisy speech has increased compared to the original noisy speech, it is expected that the fundamental frequency range is estimated more precisely in the new iteration. This in turn improves the separation performance of the system. In order to evaluate the system performance, we utilize the PESQ index. The experimental results show that after about seven iterations, the value of Δ PESQ becomes ignorable.

The block diagram of the iterative incoherent speech separation system is shown in Fig. 3. A detailed description of the iterative system follows.

As known, the performance of an adaptive filter improves as the correlation of the reference and target signals increases. Hence, to increase the quality of the reference channel signal, we preferably feed back the output of the iterative incoherent speech separation system as a new noisy input signal. This helps to achieve a target signal with a more acceptable quality in an iteration procedure.

When the output of the hybrid system is used as a new noisy input signal of the iterative incoherent speech separation system, the performance of the adaptive filter is not good enough in the first iteration. This can be justified by the low quality of the reference channel signal in the first iteration. This requires additional iterations to improve performance. Also, computational complexity of the hybrid system is high compared with the iterative incoherent speech separation system.

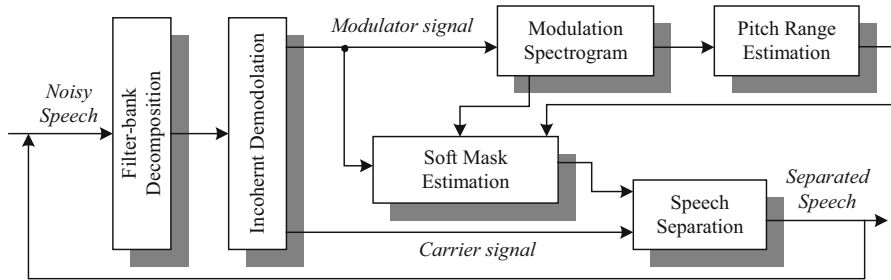


Fig. 3 Block diagram of the iterative incoherent speech separation system

2.3.1 Fundamental Frequency Range Estimation in Modulation Frequency Domain

As shown in Fig. 3, determining the fundamental frequency range of the target and interference speakers is the first step of the iterative incoherent system. This is done using an onset and offset detection algorithm [19] in the modulation frequency domain. Distribution of fundamental frequency energy in the modulation spectrogram is an important feature in determining the fundamental frequency range. Therefore, first the modulation spectrogram is calculated, and then, in the fundamental frequency range estimation stage, the modulation spectrogram intensity in each subband is smoothed over the modulation frequency using a low-pass filter. After smoothing, the modulation spectrogram is segmented using the onset and offset algorithm. Onsets and offsets associate to sudden intensity changes that can be detected from the partial derivative of smoothed modulation spectrogram intensity.

The peaks and valleys of the partial derivative of the signal are the best candidates for onsets and offsets, respectively. Those onset candidates whose peaks are greater than a threshold θ_{on} are accepted (for complete description of the method, see [27]). Between two consecutive accepted onsets, an offset with the largest intensity decrease is accepted.

Since the frequency components with close modulation frequencies of onsets or offsets probably correspond to the same source, our system connects common onsets and offsets into onset and offset fronts. An onset candidate is connected with the closest onset candidate in an adjacent subband if their distance in modulation frequency is less than a certain threshold; the same applies to the offset candidates. In the next step, segments are formed by the matching individual onset and offset fronts (for more details about the matching process, see Section 3.2.2 in [27]).

By segmentation of the modulation spectrogram of noisy speech, only segments with modulation frequencies in the range of [60,350] Hz are selected (for men, women, and children, speaker's fundamental frequency range is [60,350] Hz).

We firstly select the two segments with the least horizontal overlap in the modulation spectrogram and largest modulation spectrogram energies. These two are called segments A and B hereafter. Any other segment that is at least 80% overlapped with segment A (or B) is grouped with the segment A (or B). The remaining segments are omitted in the grouping procedure. In each subband, the range of the fundamental

frequency is determined by the onset and offset fronts of the resulting group in that subband [27].

2.3.2 Speech Separation

For a given noisy speech $x[n]$, the fundamental frequency ranges in each subband k for the target and interfering speakers are respectively denoted by $PF_{k,t}$ and $PF_{k,i}$. To produce a frequency mask for speech separation, the mean of the modulation spectral energy is defined in the fundamental frequency range as the energy normalized by the width of that fundamental frequency range for the target and interference, E_k .

By comparing the mean of modulation spectral energy of target and interference speakers (i.e., $E_{k,t}$ and $E_{k,i}$), a frequency mask is calculated:

$$F_k = \frac{E_{k,ts}}{E_{k,ts} + E_{k,is}} \quad (17)$$

To construct a filter with linear phase, the frequency mask is transformed into the time domain, which results in a filter, whose magnitude is F_k and the assigned phase is $\phi_k[i] = i$. The separated target signal is estimated by convolving the inverse DFT of the filter with the modulator signal of the mixture signal and multiplying the results by the carrier signal of the mixture signal [27].

2.4 Hybrid Coherent–Incoherent Speech Separation System

As known, in the high-frequency range of voiced speech, harmonics are generally unresolved since the corresponding auditory filters have wide passbands that respond to multiple harmonics. Psychophysical evidence suggests that the human auditory system processes resolved and unresolved harmonics differently in nonequal width subbands [31]. The bandwidth of each band-pass filter, Δf_c , is defined as:

$$\Delta f_c = 25 + 75 \times [1 + 1.4(f_c/1000)^2]^{0.69} \quad (18)$$

According to Eq. (18), the speech signal has wideband carrier frequency in high-frequency subbands, in which we have a bandwidth Δf_c instead of single carrier frequency f_c . This shows that higher harmonics of speech signals have wider bandwidths compared with lower harmonics [31]. Therefore, in the coherent demodulation method, COG, estimation of IF is problematic for wideband signals. This leads to difficulties in estimation of coherent carrier in high-frequency subbands of speech signals (see [8,26] for more details).

To solve this problem, we propose a hybrid coherent–incoherent system for single-channel speech separation. As shown in Fig. 1, the proposed system uses the coherent modulation filtering for separating low-frequency subbands of the target speaker from the interference; in addition, the high-frequency subbands of the target signal are constructed using the proposed iterative incoherent speech separation system. Therefore, the separated signal $\hat{X}_k[m]$ obtained from the proposed hybrid system for low-frequency subbands is constructed as:

$$\hat{X}_k[m] = \hat{M}_k[m] \cdot C_k^c[m] \quad k = 1, \dots, L, \quad (19)$$

and for high-frequency subbands as:

$$\hat{X}_k[m] = M_k^d[m] \cdot C_k^i[m] \quad k = L + 1, \dots, K. \quad (20)$$

where $C_k^c[m]$ and $C_k^i[m]$ are the carrier signals obtained from the coherent and incoherent demodulation methods, respectively. Experiments show that $L = 2k\text{Hz}$ is an appropriate frequency for being a boundary to separate low- and high-frequency subbands (the database used in the evaluation part is also used to compute the parameter L).

The performance of the iterative incoherent speech separation system in separating the subbands with high frequencies is superior to the coherent modulation filtering system. Thus, the signal obtained from the proposed hybrid system is superior to that obtained from the coherent modulation filtering system.

As stated, the complex modulator signal of the coherent system is closed under convolution. Also, the real modulator signal of incoherent system is able to create a modulation spectrum with high resolution and wide frequency bandwidth in the modulation frequency domain. For the speech signal, this property may be used to analyze the fundamental frequency in the modulation frequency domain [35]. Taking advantage of these two properties leads to an increase in the performance of the hybrid system.

3 Evaluation and Comparison

The proposed system is evaluated for voiced speech separation on a corpus containing 20 target utterances from the TIMIT database [13]. These utterances are mixed with different interferences (filtered through the room impulse response) at different SNR levels. The interference signals are: N0) 1-kHz pure tone, N1) white noise, N2) noise bursts (which are periodic bursts of white noise), N3) babble noise, N4) rock music, N5) siren, N6) trill telephone, N7) female speech, N8) male speech, and N9) female speech (taken from [9]) (N7, N8, and N9 are three speech signals with different fundamental frequency ranges). As shown in Table 1, these interferences are classified into three categories: 1) those with no periodicity; 2) those with quasi-periodicity; and 3) speech utterances. It is mentionable that the whole mixed (speech + noise) data are practically divided into 200-*ms* frames. Among these 200-*ms* frames, 150 frames are used for the tuning of the system parameters (especially the parameters of adaptive filtering process). The rest of these 200-*ms* frames are then used in the evaluation process.

To evaluate the proposed system, we use perceptual evaluation of speech quality (PESQ), weighted spectral slope (WSS) distance, and log likelihood ratio (LLR) as objective evaluations that correlate well with subjective mean opinion score (MOS) evaluations [18]. The value of the PESQ is limited to the interval [0,5]; higher values of PESQ indicate the superior signal. In addition, a signal with lower values of WSS and LLR is a better signal. The input signal is sampled at 16 kHz. The filterbank has 256 subbands with a prototype Hanning filter of 32-*ms* duration and a frame rate of 8

Table 1 Category of interference signals

Category 1	White noise, noise bursts
Category 2	1-kHz pure tone, babble noise, rock music, siren, trill telephone
Category 3	Female and male speech signals

Table 2 Speech separation results in different input SNRs for a mixture of the male target speaker and white noise in terms of objective measures LLR, WSS, and PESQ

SNR(dB)		-10	-5	0	5	10	15	Average
Mixture	PESQ	1.82	2.03	2.05	2.23	2.43	2.73	2.21
	WSS	42.38	34.65	29.18	26.23	21.20	16.33	28.32
	LLR	4.76	3.08	3.00	2.26	1.64	1.17	2.65
Coherent system	PESQ	2.02	2.35	2.51	2.59	2.65	2.74	2.43
	WSS	41.81	36.88	25.97	23.63	20.57	20.13	28.16
	LLR	0.71	0.67	0.62	0.60	0.59	0.58	0.62
Hybrid system	PESQ	3.12	3.43	3.75	3.79	3.93	3.97	3.63
	WSS	28.23	21.05	17.63	16.29	13.54	12.28	18.17
	LLR	0.28	0.23	0.19	0.17	0.17	0.17	0.20
Iterative system	PESQ	3.43	3.50	3.95	4.00	4.02	4.03	3.76
	WSS	20.52	18.39	17.32	15.72	13.44	9.86	15.87
	LLR	0.21	0.19	0.17	0.17	0.16	0.16	0.17

ms. The number of subbands depends on the required frequency resolution, which is determined experimentally using the speech samples from TIMIT database.

Tables 2, 3, and 4 respectively present the performance of the proposed speech separation system in terms of objective measures: PESQ, WSS, and LLR for different SNRs (in dB) before and after the separation. The results are averaged for separated target signal from the mixture of a target male speaker with a) white noise, b) babble noise, and c) male speaker, which are selected from each category listed in Table 1. The first row of each table shows PESQ, WSS, and LLR of noisy speech signal before separation. The two next rows indicate performance of the coherent system (based on modulation filtering) and the hybrid system using the incoherent speech separation system without iteration. The last row shows performance of the proposed system in which the hybrid system uses the proposed iterative incoherent speech separation system.

By comparing the results of the coherent and hybrid systems, one can clearly conclude that the performance of the hybrid system is superior to that of the coherent system. This is because the coherent system cannot separate the high-frequency subbands of the noisy speech, while the hybrid system can improve the performance of the coherent system using the incoherent speech separation system for high-frequency subbands. This is particularly noticeable at low SNRs. Also, by comparing the results of the hybrid and iterative systems, we conclude that the performance of the iterative system is superior to that of the hybrid system. Evidently, using the iterative incoher-

Table 3 Speech separation results in different input SNRs for a mixture of the male target speaker and babble noise in terms of objective measures LLR, WSS, and PESQ

SNR(dB)		-10	-5	0	5	10	15	Average
Mixture	PESQ	1.67	1.69	1.75	1.91	2.22	2.55	1.96
	WSS	59.14	48.96	39.48	34.73	28.17	19.93	38.40
	LLR	2.68	2.27	1.67	1.40	0.99	0.61	1.60
Coherent system	PESQ	0.84	1.17	2.02	2.35	2.44	2.65	1.91
	WSS	67.31	49.61	46.66	39.70	30.98	26.25	43.41
	LLR	0.75	0.69	0.68	0.63	0.62	0.60	0.66
Hybrid system	PESQ	2.57	2.84	3.20	3.49	3.70	3.87	3.27
	WSS	54.78	40.62	38.68	27.84	21.04	15.44	33.06
	LLR	1.51	0.95	0.80	0.33	0.29	0.24	0.68
Iterative system	PESQ	3.14	3.24	3.41	3.52	3.94	3.97	3.53
	WSS	40.67	33.15	23.75	18.00	14.35	10.89	23.46
	LLR	0.94	0.49	0.40	0.24	0.23	0.22	0.42

Table 4 Speech separation results in different input SNRs for a mixture of the male target and interfering speakers in terms of objective measures LLR, WSS, and PESQ

SNR(dB)		-10	-5	0	5	10	15	Average
Mixture	PESQ	0.89	1.08	1.29	1.96	2.41	3.04	1.77
	WSS	52.62	43.33	32.05	22.34	14.77	8.85	28.99
	LLR	2.22	1.57	0.97	0.61	0.37	0.18	0.98
Coherent system	PESQ	1.44	1.97	2.05	2.54	2.76	2.80	2.26
	WSS	53.90	42.57	35.52	30.86	27.67	22.40	35.48
	LLR	0.62	0.58	0.54	0.53	0.51	0.50	0.54
Hybrid system	PESQ	2.73	2.91	3.22	3.36	3.66	3.72	3.26
	WSS	37.60	33.61	23.99	23.13	17.71	12.67	24.78
	LLR	0.37	0.35	0.32	0.28	0.27	0.23	0.30
Iterative system	PESQ	3.16	3.27	3.40	3.54	3.70	3.92	3.49
	WSS	32.35	27.28	22.97	15.89	11.81	8.58	19.81
	LLR	0.36	0.33	0.30	0.23	0.23	0.22	0.27

ent system for estimating the reference channel for coherent filtering is the key to the performance improvement.

As an additional benchmark, we have evaluated the proposed system on a corpus of 100 mixtures composed of 10 target utterances mixed with 10 intrusions described in the beginning of Sect. 3 [9]. This corpus is commonly used for CASA research [7, 11, 42]. The database contains utterances from both male and female speakers. In these experiments, we evaluate the performance of speech separation systems in terms of PESQ, WSS, and LLR. Also, we employ three other measures, namely percentage of energy loss (P_{EL}), percentage of residual noise (P_{NR}), and SNR [20].

P_{EL} measures the amount of the target speech excluded from the segregated speech as:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n s^2(n)}, \quad (21)$$

P_{NR} measures the amount of intrusion remaining in the segregated speech,

$$P_{EL} = \frac{\sum_n e_2^2(n)}{\sum_n \tilde{x}^2(n)}, \quad (22)$$

and SNR of the segregated target (in dB) provides a good comparison between the waveforms, which is calculated as:

$$SNR = 10 \log_{10} \frac{\sum_n s^2(n)}{\sum_n [s(n) - \tilde{x}(n)]^2}. \quad (23)$$

where $\tilde{x}(n)$ is the estimated signal and $s(n)$ is the target signal before mixing with intrusion. $e_1(n)$ denotes the signal present in $s(n)$ but absent in $\tilde{x}(n)$, and $e_2(n)$ denotes the signal present in $\tilde{x}(n)$ but absent in $s(n)$.

Figure 4 shows the average SNR in the case of each intrusion for the output of the proposed system (the iterative system) compared with those for the original mixtures, Hu and Wang system 2004 (HW'04) [20], Hu and Wang system 2010 (HW'10) [21], and our pervious system, which is called the “incoherent system” [27], hereafter. The average SNR of our proposed system, taken over all input SNR levels, is 18.73 dB. All four systems show improvements compared to the original mixtures. The average SNR improvement of the proposed system for the entire corpus compared with the

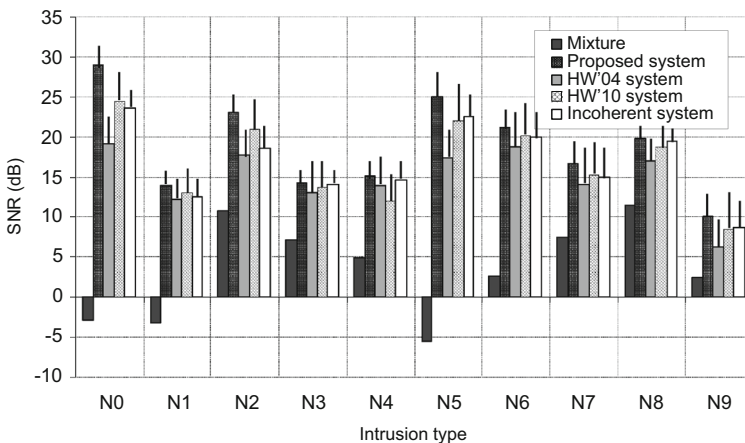


Fig. 4 SNR results for segregated speech and original mixtures for a corpus of voiced speech and various intrusions

Table 5 Speech separation results for different methods in terms of objective measures LLR, WSS, and PESQ for each intrusion (N0–N9) mixed with 10 target utterances

Intrusion		N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Average
Mixture	PESQ	1.87	1.30	1.63	1.54	1.60	0.30	1.79	1.77	1.69	1.56	1.50
	WSS	81.98	39.76	17.04	33.03	34.37	62.94	65.25	40.31	20.19	36.60	43.14
	LLR	0.44	4.14	1.22	1.44	1.63	1.80	1.05	1.05	0.83	1.50	1.51
Proposed system	PESQ	3.74	3.68	3.77	3.59	3.71	3.69	3.79	3.62	3.76	3.49	3.68
	WSS	7.58	17.95	7.77	14.56	9.90	18.41	7.37	10.86	9.66	16.21	12.02
	LLR	0.22	0.19	0.22	0.24	0.21	0.29	0.22	0.23	0.23	0.32	0.23
HW'04 system	PESQ	2.84	1.44	2.13	1.61	1.39	2.42	2.61	1.73	2.06	1.21	1.94
	WSS	35.91	59.01	26.86	50.52	62.86	34.49	40.20	46.26	30.38	64.21	45.07
	LLR	5.51	2.87	4.75	3.76	3.23	4.30	3.89	4.94	4.72	4.77	4.27
HW'10 system	PESQ	2.64	1.24	2.14	1.30	1.14	2.21	2.24	1.62	1.78	1.65	1.79
	WSS	35.31	46.36	25.33	48.24	55.19	41.71	34.33	38.77	24.83	51.82	40.18
	LLR	5.70	3.03	4.53	3.92	3.98	5.37	3.93	4.00	4.50	4.78	4.37
Incoherent system	PESQ	3.17	2.82	3.10	2.69	2.98	3.02	3.06	2.85	3.04	2.75	2.94
	WSS	20.61	28.97	12.02	24.68	17.08	30.18	14.13	29.89	20.85	34.07	23.24
	LLR	0.45	0.53	0.39	0.64	0.50	0.72	0.43	0.54	0.64	1.43	0.62

HW'10 system is about 1.86 dB. The HW'10 system in turn performs 1.96 dB better on average than HW'04 system. In addition, the proposed system shows 1.88 dB SNR improvement over the incoherent system. Larger SNR improvements are obtained for intrusions whose modulation spectra do not significantly overlap with those of the target utterances (e.g., N0 and N5), whereas improvements are modest for intrusions with significant overlap (e.g., N3 and N8). To analyze the significance of the results, the p-value is calculated in the case of each noise for the results presented in Fig. 4 (i.e., the proposed system, HW'04, HW'10, and our pervious system). The results show that the p-value in the case of N0, N2, N5, and N8 is very small which indicates that the differences between the average performances of different separation systems are highly significant. On the other hand, the p-value in the case of N1 and N3 is close to 0.7, in N6 case is close to 0.5, and in the case of N4, N7, and N9 is close to 0.1.

Table 5 shows performance of the proposed, HW'04, HW'10, and the incoherent systems in terms of objective measures of PESQ, WSS, and LLR for each intrusion (N0–N9) mixed with the target speech at different SNRs (as presented in Fig. 4). Several results are inferred from Table 5: 1) According to the table, the proposed system achieves a higher PESQ score compared with other systems; 2) it is observed that the proposed system achieves about one unit improvement in PESQ score over both of the HW's systems for N0, N2, N5, and N6 intrusions and two units improvement for other intrusions; and 3) the proposed system outperforms the other ones, as it is evident from lower LLR and WSS scores.

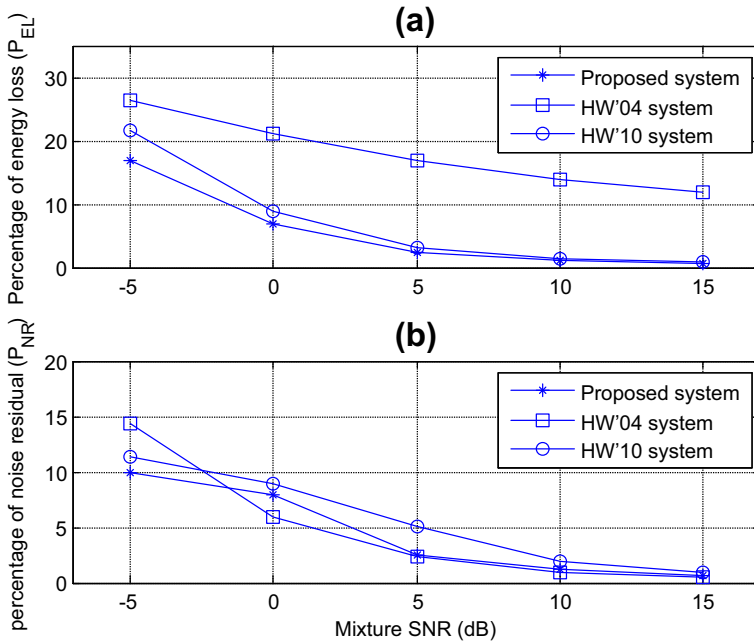


Fig. 5 Results of voiced speech separation. (a) Percentage of energy loss on voiced target. (b) Percentage of noise residue

The ranges of LLR and WSS of the proposed system are (0.19–0.32) and (7.37–18.41), while those for the HW's systems are (2.87–5.70) and (24.83–64.21), respectively. Comparing the results depicted in Fig. 4 and Table 5, it can be concluded that although the SNR obtained from the HW'10 system is higher than that of their previous method (HW'04), their PESQ scores are almost the same. This means that although the HW'10 system achieves more noise elimination, it leads to more target signal cancellation. In contrast, our proposed method has concurrently improved SNR and PESQ.

Figure 5a, b shows the average values for percentage of energy loss, P_{EL} , and noise residue, P_{NR} , at a particular SNR level. By comparing the depicted results, it is concluded that the energy loss percentage of the proposed system presents a significant improvement over those of HW's systems. At the same time, the percentage of noise residue is lower than those of HW'10 system for all SNR's and HW'04 system for -5 dB SNR.

As shown in Fig. 5, the proposed system segregates 93% of the voiced target energy at 0 dB SNR and 99.3% at 15 dB SNR. At the same time, at 0 dB, 8% of the segregated energy belongs to the intrusion. This number drops to 0.7% at 15 dB SNR. Although the HW'04 system yields a lower percentage of noise residues (6% at 0 dB and 0.5% at 15 dB, as Fig. 5b reveals), its percentages of target energy loss are much higher (78.8% at 0 dB and 88% at 15 dB, see Fig. 5a). It is concluded that the proposed system reduces the signal distortion but with a slight increase in P_{NR} .

To help the reader verify the actual differences in the performance, a file has been prepared including sample audio mixture signals (target speech + interfer-

ence) and the results of the separation using the proposed system, HW's systems, and the incoherent system. The file is available at <http://pws.yazd.ac.ir/sprl/Mahmoodzadeh-CSSP-SampleWaves.ppt>.

4 Conclusion

This paper tackles the issue of single-channel speech separation, which is a very challenging problem. We propose a hybrid system of coherent and incoherent speech separation methods for separation of target speech from the interference. In our approach, the coherent method is used for the speech separation based on adaptive modulation filtering. In this method, first the COG method decomposes the subband speech signal into the modulator and carrier components for the coherent demodulation method. Then, by using the adaptive AP filtering on each subband of the modulator signal, the target speech is separated from the interference signal.

Our system applies an iterative incoherent speech separation system that simultaneously provides the reference channel for the adaptive coherent modulation filtering. This method separates speech from the interference using a soft mask obtained in the modulation spectrum domain. For this purpose, first the fundamental frequency ranges of target and interference are determined in modulation spectrogram domain using an onset and offset algorithm. Then, the soft mask is determined considering the fundamental frequency range estimated in each subband and the modulation spectrum energy of the target speech in that range. Also, considering the difficulties of coherent carrier estimation in high-frequency subbands, we propose a system that employs coherent modulation filtering for low-frequency subbands and incoherent modulation filtering for high-frequency subbands.

In terms of computational complexity, the main extra cost of the proposed hybrid system (compared to that of pervious incoherent speech separation system) arises from the adaptive filter used in the adaptive coherent modulation filtering system. Since we have employed the AP algorithm, whose computational complexity is close to LMS algorithm, the excess computational complexity is ignorable. The time period for the iteration and the processing of speech separation system is short.

It should be noted that the proposed system can also separate the quasi-periodic unvoiced portions due to their proximity with the voiced parts; also, in practice, the inability in separating the unvoiced portions does not have a large impact on the performance of the system.

Evaluations and comparisons based on standard measures show that the proposed hybrid system produces substantial SNR gains across different interferences and outperforms similar separation methods.

References

1. H.R. Abutalebi, H. Sheikhzadeh, R.L. Brennan, G.H. Freeman, Affine projection algorithm for over-sampled subband adaptive filters. in Proceedings of IEEE ICASSP (2003), pp. 209–212
2. T. Arai, M. Pavel, H. Hermansky, C. Avendano, Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelopes. in Proceedings of ICSLP (1996), pp. 90–93

3. L. Atlas, P. Clark, S. Schimmel, Modulation Toolbox Version 2.1 for MATLAB. (2012). <https://sites.google.com/a/uw.edu/isdl/projects/modulation-toolbox>
4. L. Atlas, C. Janssen, Coherent modulation spectral filtering for single-channel music source separation. in Proceedings of IEEE ICASSP (2005), vol. IV, pp. 461–464
5. L. Atlas, Q. Li, J. Thompson, Homomorphic modulation spectra. in Proceedings of IEEE ICASSP (2004), vol. II, pp. 761–764
6. L. Benaroya, F. Bimbot, R. Gribonval, Audio source separation with a single sensor. IEEE Trans. Audio Speech Lang. Process. **14**(1), 191–199 (2006)
7. G.J. Brown, M. Cooke, Computational auditory scene analysis. Comput. Speech Lang. **8**, 297–336 (1994)
8. P. Clark, L. Atlas, Time-frequency coherent modulation filtering of non-stationary signals. IEEE Trans. Signal Process. **57**(11), 4323–4332 (2009)
9. M.P. Cooke, *Modeling Auditory Processing and Organization* (Cambridge Univ. Press, Cambridge, 1993)
10. M. Cooke, J.R. Hershey, S.J. Rennie, Monaural speech separation and recognition challenge. Comput. Speech Lang. **24**(1), 1–15 (2010)
11. L.A. Drake, Sound source separation via computational auditory scene analysis (CASA) enhanced beamforming. Ph.D. dissertation, Dept. Elect. Eng., Northwestern Univ., Evanston, IL, 2001
12. R. Drullman, J. Festen, R. Plomp, Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. **95**(2), 1053–1064 (1994)
13. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus. (1993). <http://www ldc.upenn.edu/Catalog/LDC93S1.html>
14. O. Ghizta, On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J. Acoust. Soc. Am. **110**(3), 1628–1640 (2001)
15. S. Greenberg, B.E.D. Kingsbury, The Modulation Spectrogram: in Pursuit of an Invariant Representation of Speech,” in Proceedings of IEEE ICASSP, (1997) pp. 1647–1650
16. L. Gu, R.M. Stern, Single-channel speech separation based on modulation frequency. in Proceedings of IEEE ICASSP (2008), pp. 25–28
17. J. Hershey, S. Rennie, P. Olsen, T. Kristjansson, Super-human multitalker speech recognition: a graphical modeling approach. Comput. Speech Lang. **24**(1), 45–66 (2010)
18. Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2008)
19. G. Hu, D.L. Wang, Auditory segmentation based on onset and offset analysis. IEEE Trans. Audio Speech Lang. Process. **15**(2), 396–405 (2007)
20. G. Hu, D.L. Wang, Monaural speech separation based on fundamental frequency tracking and amplitude modulation. IEEE Trans. Neural Netw. **15**, 1135–1150 (2004)
21. G. Hu, D.L. Wang, A tandem algorithm for fundamental frequency estimation and voiced speech segregation. IEEE Trans. Audio Speech Lang. Process. **18**(8), 2067–2079 (2010)
22. M. Karjalainen, T. Tolonen, Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. in Proceedings of IEEE ICASSP (1999), pp. 929–932
23. A. Kusumoto, T. Arai, T. Kitamura, M. Takahasi, Y. Murahara, Modulation enhancement of speech as preprocessing for reverberant chambers with the hearing-impaired. in Proceedings of IEEE ICASSP (2000), pp. 853–856
24. H. Laurberg, Uniqueness of non-negative matrix factorization. in Proceedings of IEEE 14th Workshop on Statistical Signal Processing (2007), pp. 44–48
25. P. Li, Y. Guan, S. Wang, B. Xu, W. Liu, Monaural speech separation based on MAXVQ and CASA for robust speech recognition. Comput. Speech Lang. **24**(1), 30–44 (2010)
26. P.J. Loughlin, B. Tacer, On the amplitude- and frequency-modulation decomposition of signals. J. Acoust. Soc. Am. **100**(3), 1594–1601 (1996)
27. A. Mahmoodzadeh, H.R. Abutalebi, H. Soltanian-Zadeh, H. Sheikhzadeh, Single channel speech separation in modulation frequency domain based on a novel fundamental frequency range estimation method. EURASIP J. Adv. Signal Process. (2012). doi:10.1186/1687-6180-2012-67
28. A. Mahmoodzadeh, H. Sheikhzadeh, H.R. Abutalebi, H. Soltanian-Zadeh, A hybrid coherent-incoherent method of modulation filtering for single channel speech separation. in Proceedings of IEEE ICASSP (2012), pp. 329–332

29. J. Ming, T. Hazen, J. Glass, Combining missing-feature theory, speech enhancement, and speaker-dependent-independent modeling for speech separation. *Comput. Speech Lang.* **24**(1), 67–76 (2010)
30. P. Mowlae, New Strategies for Single-channel Speech Separation. Ph.D. dissertation, Dept. Electronic Systems, Aalborg University (2010)
31. L.R. Rabiner, R.W. Schafer, *Theory and Applications of Digital Speech Processing* (Prentice-Hall Inc, Upper Saddle River, 2011)
32. M. Radfar, R. Dansereau, A. Sayadiyan, Monaural speech segregation based on fusion of source-driven with model-driven techniques. *Speech Commun.* **49**(6), 464–476 (2007)
33. S.T. Roweis, One microphone source separation. in *Proceedings of Neural Information Processing Systems* (2000), pp. 793–799
34. S.G. Sankaran, A.A. Beex, Convergence behavior of affine projection algorithms. *IEEE Trans. Signal Process.* **48**(4), 1086–1096 (2000)
35. S.M. Schimmel, Theory of modulation frequency analysis and modulation filtering with applications to hearing devices. Ph.D. dissertation, Dept. Comput. Sci. & Eng., Washington University (2007)
36. S.M. Schimmel, L.E. Atlas, Coherent envelope detection for modulation filtering of speech. in *Proceedings of IEEE ICASSP* (2005), vol. I, pp. 221–224
37. S.M. Schimmel, L.E. Atlas, K. Nie, Feasibility of single channel speaker separation based on modulation frequency analysis. in *Proceedings of IEEE ICASSP* (2007), vol. 4, Apr., pp. 605–608
38. S.M. Schimmel, K.R. Fitz, L.E. Atlas, Frequency reassignment for coherent modulation filtering. in *Proceedings of IEEE ICASSP* (2006), vol. 5, pp. 261–264
39. Y. Shao, S. Srinivasan, Z. Jin, D.L. Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition. *Comput. Speech Lang.* **24**(1), 77–93 (2010)
40. M. S. Vinton, L. E. Atlas, Scalable and progressive audio codec. in *Proceedings of IEEE ICASSP* (2001), vol. 5, pp. 3277–3280
41. D.L. Wang, G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE Press, Hoboken, 2006)
42. D.L. Wang, G.J. Brown, Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Netw.* **10**, 684–697 (1999)
43. M. Wu, D.L. Wang, G.J. Brown, A multi-pitch tracking algorithm for noisy speech. in *Proceedings of IEEE ICASSP* (2002), pp. 369–372