EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Single channel speech separation in modulation frequency domain based on a novel pitch range estimation method

Azar Mahmoodzadeh[1], Hamid Reza Abutalebi[1*], Hamid Soltanian-Zadeh[2,3] and Hamid Sheikhzadeh[4]

Abstract

Computational Auditory Scene Analysis (CASA) has been the focus in recent literature for speech separation from monaural mixtures. The performance of current CASA systems on voiced speech separation strictly depends on the robustness of the algorithm used for pitch frequency estimation. We propose a new system that estimates pitch (frequency) range of a target utterance and separates voiced portions of target speech. The algorithm, first, estimates the pitch range of target speech in each frame of data in the modulation frequency domain, and then, uses the estimated pitch range for segregating the target speech. The method of pitch range estimation is based on an onset and offset algorithm. Speech separation is performed by filtering the mixture signal with a mask extracted from the modulation spectrogram. A systematic evaluation shows that the proposed system extracts the majority of target speech signal with minimal interference and outperforms previous systems in both pitch extraction and voiced speech separation.

Keywords: acoustic frequency, modulation frequency, onset and offset algorithm, pitch range estimation, speech separation

1. Introduction

Speech separation, as a solution to the cocktail party problem, is a well-known challenge with important applications. To touch the point, consider the telecommunication systems or the Automatic Speech Recognition systems that lose performance in the presence of interfering sounds [1,2]. An effective system that segregates speech from interference in monaural (single-microphone) situations can be rewarding in such problems. Many methods have been proposed for monaural speech enhancement; for example, see [3-7]. These methods usually assume certain statistical properties for interference and tend to lack the capacity of dealing with a variety of interferences. While the monaural speech separation works awkwardly, the human auditory system performs proficiently. The perceptual process is considered as Auditory Scene Analysis (ASA) [5]. Psychoacoustic research in ASA has inspired considerable work in developing Computational Auditory Scene Analysis (CASA) systems for speech separation (see [6,7] for a comprehensive review).

According to Bregman [5], ASA procedure can be separated into two theoretical stages: *segmentation* and *grouping*. At the first stage, speech is transformed into a higher-dimensional space (such as a time-frequency two-dimensional representation) and then, similar time-frequency (T-F) units are segmented in order to compose different regions [6]. In the second stage, these regions are combined into different streams based on the relevant acoustic information. The major computational goal of CASA is to separate the target speech signal from the interference for different purposes, via generating a binary or a soft T-F mask, see, e.g., [8-10].

Grouping, itself, consists of simultaneous and sequential organizations, which involves grouping of segments across frequency and time. The task of sequential grouping is to group the T-F regions relative to the same sound source across time. Figure 1 illustrates this issue in which the upper panel shows T-F regions grouped into one single stream, as they are close enough in both (time

* Correspondence: habutalebi@yazduni.ac.ir
[1]Speech Processing Research Lab (SPRL), Electrical and Computer Engineering Department, Yazd University, Yazd, Iran
Full list of author information is available at the end of the article

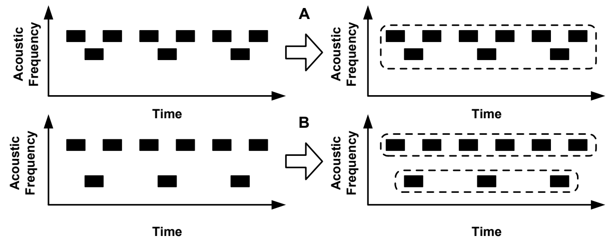

Figure 1 Segmentation and grouping of speech projected into T-F cells in a 2D representation [6].

and frequency) directions; while, the lower panel illustrates the case of two streams of speech, grouped separately as the T-F regions are sufficiently far from each other in the frequency direction. Temporal continuity is an effective cue for grouping T-F regions neighboring in time. However, it cannot handle T-F regions that do not overlap in time due to the silence or interference segments. Therefore, sequential grouping of such T-F regions is a very challenging problem (see [11,12] for more details).

Natural speech includes both voiced and unvoiced portions. Voiced portions of speech are described by *periodicity* (or *harmonicity*), which has been used as an important feature in many CASA systems for segregating voiced speech (see, e.g. [13,14]). Despite considerable advances in voiced speech separation, the performance of current CASA systems is still limited by pitch frequency (F0) estimation errors and residual noise. Various methods have been proposed for robust pitch frequency estimation, see e.g., [15,16]; however, robust pitch frequency estimation in low signal-to-noise ratio (SNR) situations still poses a significant challenge.

While mixed speech may have a great deal of overlap in the time domain, modulation frequency analysis provides an additional dimension that can present a greater degree of separation among sources. In other words, the original T-F representation obtained from transformations like Short-Time Fourier Transform (STFT) can be augmented to a third dimension that represents modulation frequency. In [17], by assuming that the pitch frequency range is known and this range is constant in each filter channel, the modulation spectral analysis is used as a tool for producing the mask for speech separation a higher-dimensional spaces.

Based on the above observations, we propose a new system for single channel separation of voiced speech based on the modulation filtering. The idea is that, first, the target pitch (frequency) range is estimated in the modulation frequency domain, and then, this range is used for producing the proper mask for speech separation. Because of the following reasons provided in [18], modulation analysis and filtering are applied for the target speech separation problem. First, there is a general belief stating that the human ASA system processes the sounds in the modulation frequency domain. Second, the energy from two co-channel talkers is largely non-overlapping in the modulation frequency domain. The method of modulation analysis and filtering has extensively been studied by many researchers in the field of single channel speech separation; Reference [19] provides a general discussion on this subject.

At first, the proposed system performs a multipitch range estimation of target and interference speech based on the segmentation of modulation spectrogram domain. The segmentation is done using an onset and offset algorithm similar to that proposed by Hu and Wang [20]. In the proposed method, the noisy signal is divided into 200 ms time frames and then, the proposed speech separation algorithm is applied to each individual frame. Pitch range estimation method works in three stages: the first stage computes the modulation spectrogram; the second stage decomposes the modulation spectrogram into segments using an onset and offset algorithm. In this stage, at first, the peaks and valleys of derivative smoothed intensity of modulation spectrogram are detected and marked as onset and offset candidates. Any onset bigger than a certain threshold is accepted for which the smallest offset between two onsets is selected. Then, onset and offset fronts are produced by connecting the common onsets and offsets. Finally, the segments are formed by matching the onset and offset fronts. The third stage determines the range of pitch frequency by selecting and grouping the desired segments.

The separation part of the proposed system aims at obtaining a soft mask in the modulation spectrogram domain. By extending the soft mask suggested in [17], a soft mask is proposed whose value depends on the estimated pitch range in each filter channel. To determine the soft mask in each filter channel, first, we find and mutually compare the modulation spectrogram energy of target and interference in their pitch ranges estimated from the previous stage. Then, we transform the soft mask to the time domain and filter the mixture signal in order to obtain the separated target signal. Thus, a strategy is suggested which estimates the target pitch range, and subsequently, segregates the target signal from the interference. Finally, the separated target signal is obtained from arranging the separated signal from each frame, in a time order sequence.

This article is organized as follows. Section 2 describes the modulation frequency analysis. In Section 3, first, a brief description of the present system is given and then the details of each stage are presented. In Section 4, a quantitative measure is proposed for evaluating the performance of speech separation and it is used for systematic evaluation of pitch range estimation and speech separation. This article concludes with a discussion in Section 5.

2. Modulation frequency analysis

Decomposing a narrowband signal into a carrier and a modulator signals is an important problem in modulation analysis and filtering [18]. The modulator is a low-frequency signal that describes the amplitude modulation of the original signal; and the carrier is a narrowband signal describing the frequency modulation of the signal. Consider a wideband discrete-time signal $x(n)$, for which n represents a discrete-time independent variable. The T-F transform of a signal $x(n)$, denoted by $X(m, k)$, is obtained using the Discrete STFT (DSTFT). $X(m, k)$ is a T-F transformed narrowband signal (with the time index m) coming out of the kth channel:

$$\mathrm{DSTFT}\left\{x(n)\right\} = X(m,k) = \sum_{n=0}^{K-1} x(n)\, w(m\mathrm{M} - n)\, e^{-j2\pi nk/K} \quad k = 0, \ldots, K-1, \tag{1}$$

where K is the DSTFT length (equal to the number of the filter bank channels), $w(\cdot)$is the acoustic frequency analysis window with length L and M is the decimated factor. The product model of the modulator signal $M(m, k)$ and the carrier signal $C(m, k)$ of the signal $X(m, k)$ in the T-F domain is defined as

$$X(m,k) = M(m,k)\;\; C(m,k), \tag{2}$$

The modulator of the signal $X(m, k)$ is found by applying an envelope detector to this signal, as

$$M(m,k) \triangleq \mathrm{D}\left\{X(m,k)\right\}, \tag{3}$$

where D is the operator of the envelope detector. With respect to Equation (2), the signal's carrier is described as

$$C(m,k) = \frac{X(m,k)}{M(m,k)}, \tag{4}$$

A good choice for the envelope detector is the *incoherent detector*, since it is able to create a modulation spectrum that has a large area covered in the modulation frequency domain. For the speech signal in hand, this property may be used to find the pitch frequency in the modulation frequency domain. Incoherent envelope detector is based on the Hilbert envelope (for real-valued subbands) or the magnitude operator (for complex-valued subbands) [21]. Therefore, the modulator of the complex signal $X(m, k)$ is defined as

$$M(m,k) = \left|X(m,k)\right|, \tag{5}$$

The theory of modulation frequency analysis and filtering is best explained through the definition of modulation transforms, which are signal transformations defined based on the Fourier transform (FT) and the STFT. The discrete short-time modulation transform of the signal $x(n)$ is defined as

$$\begin{aligned} X(k, i) &= \mathrm{DFT}\left\{\mathrm{D}\left\{\mathrm{DSTFT}\left\{x(n)\right\}\right\}\right\} \\ &= \sum_{m=0}^{I-1} M(m,k)\, e^{-j2\pi mi/I} \quad i = 0, \ldots, I-1, \end{aligned} \tag{6}$$

where I is the DFT length and i is the modulation frequency index. The modulation transform consists of a filter-bank that uses the DSTFT followed by a subband envelope detector and, then, a frequency analyzer of the subband envelopes (the DFT) [18].

The modulation spectrogram intensity, defined as $\mathcal{X}(k,i) = \left|X(k,i)\right|$, is generally sketched in a diagram, in which the vertical axis displays the regular acoustic frequency index k and the horizontal axis is the modulation frequency index i. The modulation analysis framework is described in Figure 2. A typical example of modulation transform is illustrated in Figure 3, in which, Figure 3a shows the mixture of a target and interfering male speakers and Figure 3b, c, respectively, depict the corresponding T-F representation and modulation spectrogram, with the overall SNR of 0 dB.

3. System description

The main target of the current system is to produce a soft mask for single channel speech separation in the modulation spectrogram domain. In the proposed system, determining the pitch range of target and interference speech is necessary for producing the mask for speech separation. The value of this mask in each subband depends on the obtained pitch range of target and interference in that

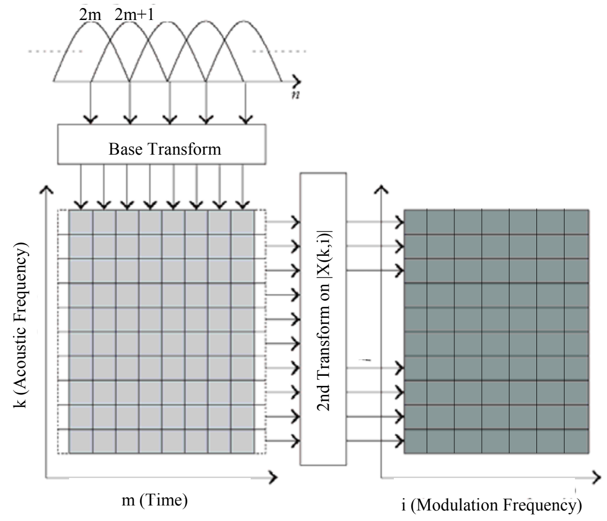

Figure 2 The modulation analysis framework and the modulation spectrogram [19].

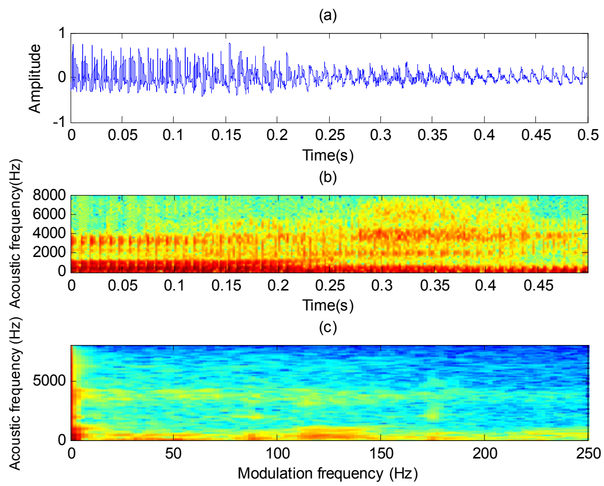

Figure 3 Sound mixture and its modulation spectrogram. **(a)** Mixture of speech signals. **(b)** T-F energy plot for a mixture of two utterances of a male speaker. The utterances are "eight" and "dos." For better display, energy is plotted as the square of the FT. **(c)** Modulation spectrogram of the mixture signal.

subband. When the modulation spectrogram of the speech signal is computed, the pitch ranges of target and interference speakers are determined and, then, a proper mask is calculated for the speech separation. The overall stages of our system are shown in Figure 4.

To determine the mentioned pitch ranges, our proposed method uses an onset and offset detection algorithm [20] to find the distribution of modulation spectrogram energy in the modulation frequency domain, which is an important feature for determining the pitch range. When modulation spectrogram energy is found, the modulation spectrogram is segmented, as described in Section 3.2.2. Then, the resulting segments are grouped in order to estimate the pitch range of each speaker. A detailed description of stages is as follows.

3.1. T-F decomposition and modulation transform

At the T-F stage, the STFT (as a uniform filter-bank) is used for decomposing a broadband signal into narrowband subband signals. The output of the T-F stage enters into the modulation transform stage in order to calculate the modulation spectrogram.

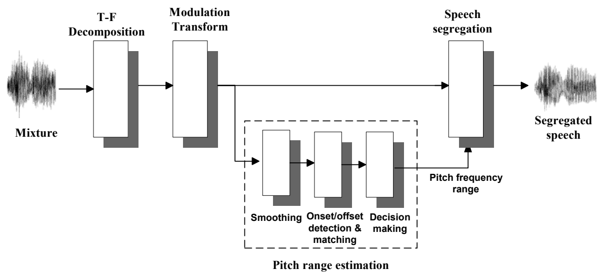

Figure 4 Block diagram of the proposed system.

3.2. Pitch range estimation in modulation frequency domain

The pitch frequencies of target and interference speakers are both time-varying. Occasionally, pitch frequencies of the target and interference speakers are too close to each other, in which this fact causes undesired errors in multipitch tracking algorithms and decreases the accuracy of speech separation methods. The algorithm of this article estimates the pitch range of target and interference speakers of noisy speech in the modulation frequency domain. Estimating the pitch range in small time-intervals (for example 200 ms) decreases the error in the pitch range estimation method.

In the pitch range estimation approach, at first, the intensity of the modulation spectrogram is smoothed over the modulation frequency, using a low-pass filter. Then, the partial derivative of the smoothed intensity over the modulation frequency is computed. By marking the peaks and valleys of the resulting signal, the onset and offset candidates are detected and the onset and offset fronts are formed. By matching the onset and offset fronts, the modulation spectrogram of speech signal is segmented. The detailed description of the stages for the pitch range estimation is as follows.

3.2.1. Smoothing

Smoothing corresponds to low-pass filtering. The proposed system uses a low-pass filter to smooth the modulation spectrogram intensity over the modulation frequency. Considering the frequency channel k, the smoothed intensity for $\mathcal{X}(k,i)$ is found as follows:

$$\mathcal{X}_s(k,i) = \mathcal{X}(k,i) * g_s(i), \tag{7}$$

where $g_s(i)$ is a low-pass FIR filter with a small number of coefficients with pass-band [0, s] in Hz. Here, "*" denotes the convolution operator (over the modulation frequency). The parameter s determines the degree of smoothing: the smaller s, the smoother $\mathcal{X}_s(k,i)$ would be.

As an example, Figure 5 shows the original (Figure 5a) and the smoothed (Figure 5b-d) intensities of the modulation spectrum for the mixture input signal shown in Figure 3a, at three typical scales. To display more details, Figure 5e-h describes the original and the smoothed intensities at these three scales, in a single frequency channel centered at 560 Hz. The intensity fluctuation reduces by smoothing, as certified by Figure 5. Although the local details of onsets and offsets become blurred, the major intensity changes of the onsets and offsets are still preserved.

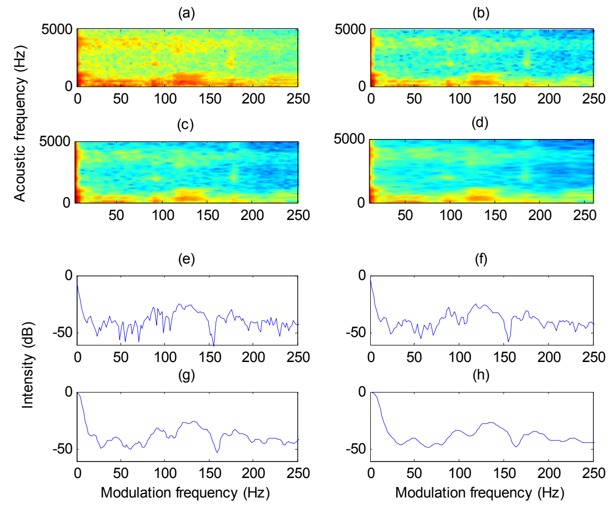

Figure 5 Smoothed intensity values at different scales. (a) Initial intensity for all channels. **(b)** Smoothed intensity at the scale 14. **(c)** Smoothed intensity at the scale 10. **(d)** Smoothed intensity at the scale 4. **(e)** Initial intensity in a channel centered at 560 Hz. **(f)** Smoothed intensity in the channel at the scale 14. **(g)** Smoothed intensity in the channel at the scale 10. **(h)** Smoothed intensity in the channel at the scale 4. The input is shown in Figure 3a.

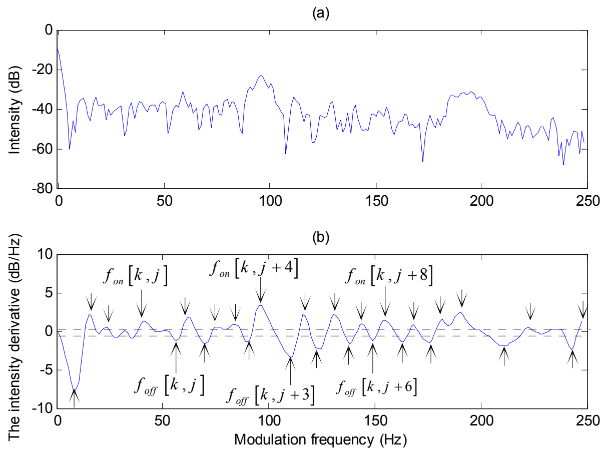

Figure 6 Onset and offset detection. The upper panel shows the response intensity and the lower panel shows the results of onset and offset detection using a low-pass filter. The threshold for onset detection is 0.05 and for offset detection is -0.05 indicated by the dash-lines. Detected onsets are marked by downward arrows, and offsets by upward arrows.

3.2.2. Onset/offset detection and matching

Onsets and offsets correspond to sudden intensity changes. The partial derivative of smoothed modulation spectrogram intensity over the modulation frequency is obtained as

$$\frac{\partial}{\partial i}\mathcal{X}_s(k,i) = \frac{\partial}{\partial i}\left[\mathcal{X}(k,i) * g_s(i)\right], \tag{8}$$

Peaks and valleys of the resulting signal of Equation (8) are, respectively, marked as onset and offset candidates. Figure 6 illustrates this procedure, in which the onset candidates with peaks bigger than a threshold θ_{on} are accepted. The peaks corresponding to the true onsets are usually significantly higher than other peaks. For this reason, $\theta_{\text{on}} = \mu + \sigma$ is selected as the threshold, in which μ and σ are the mean and standard deviation of all the onset candidates (peaks of Equation 8), respectively [20]. Hu and Wang [20] claim that the performance of the method using such a threshold choice is satisfactory.

In every filter channel k, to determine the offset corresponding to each onset candidate, let $f_{\text{on}}[k, l]$ represent the modulation frequency of the lth onset candidate in the filter channel k. The corresponding offset, denoted by $f_{\text{off}}[k, l]$, is located between $f_{\text{on}}[k, l]$ and $f_{\text{on}}[k, l+1]$. If there are multiple offset candidates in this interval, the one with the largest intensity decrease (i.e., the smallest $\frac{\partial}{\partial i}\mathcal{X}_s(k,i)$) is chosen.

After finding the onsets and offsets, those with close modulation frequencies are connected to the onset and offset fronts, because the frequency components of onsets and offsets with close modulation frequencies probably correspond to the same source. Onset and offset fronts are vertical contours across acoustic frequency in the modulation spectrogram domain. The proposed system connects an onset candidate from a filter channel to an onset candidate in the above adjacent filter channel, provided that their distance in the modulation frequency is less than a certain threshold relative to the latter filter channel. In each filter channel, this threshold is defined as the mean of the distances in the modulation frequency direction between two-by-two adjacent onsets. This definition for the threshold is provided from experiments and is validated as a good choice in the data. The same applies to the offset candidates. Notice that a threshold with a too small value may prevent onsets or offsets from the same event to joint; while a threshold with a too large value may cause some onsets from different events to connect together [20].

The next step is to form segments by matching individual onset and offset fronts. Consider ($f_{\text{on}}[k, l_k]$, $f_{\text{on}}[k, l_{k+1}]$,..., $f_{\text{on}}[k+r\text{-}1, l_{k+r\text{-}1}]$) as an onset front with r consecutive filter channels, in which l_k denotes the number of the selected onset as an onset front member, in the filter channel k; and consider ($f_{\text{off}}[k, l_k]$, $f_{\text{off}}[k+1, l_{k+1}]$,..., $f_{\text{off}}[k+r\text{-}1, l_{k+r\text{-}1}]$) as the corresponding offset modulation frequencies. For each offset modulation frequency, first, we find all those offset fronts that cross this offset; then, the offset front with the most crosses (with the offset modulation frequencies) is chosen as the matching offset front. Now, the entire filter channels from k to $k + r - 1$ occupied by the matching offset front (and their corresponding offset modulation frequencies on this matching offset front) are labeled as "matched." If all the channels from k to $k+r-1$ are labeled as matched, the matching procedure finishes; otherwise, the matched channels should be put aside and the procedure should be repeated for the remaining unmatched channels.

At last, in order to form the offset front relative to each onset front, we replace the offset modulation frequencies corresponding to the onset front with those of the matched offset fronts. The region between the onset front and its offset front yields a 2D segment in the acoustic-modulation frequency space; see Figure 7 for the schematic representation of the matching procedure.

3.2.3. Segment selection and decision-making

By detecting the onsets and offsets and forming the onset and offset fronts, the modulation spectrogram domain of speech signal is segmented. Since the speaker's pitch range is [60, 350] Hz (for men, women, and children), only the segments with modulation frequencies in this range are accepted. Now, we describe the grouping procedure for the segments.

First, the modulation spectrogram energy of each selected segment is computed. Two almost disjoint segments with most energies, i.e., those with the most modulation spectrogram energies and the least horizontal overlap in the modulation spectrogram, for simplicity called segments A and B, are selected (the case "speech interfered by a non-speaker-noise" has only one such segment). For any other segment (call segment C), if the modulation frequency range at least 80% overlaps with that of segment A or segment B, the segment C is grouped with that overlapping segment; otherwise, the segment C is omitted for the grouping procedure. Figure 8 presents a typical example of the grouping procedure. As shown, in each filter channel, the onset and offset fronts of the resulting group determines the corresponding range of pitch frequency in that filter channel.

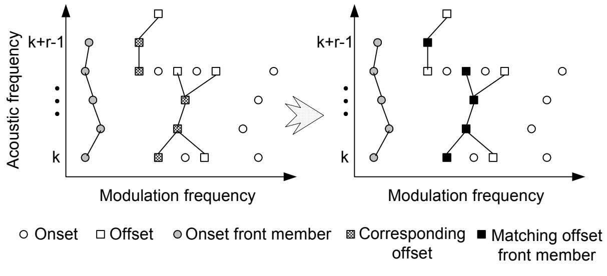

Figure 7 Schematic representation of the matching procedure; (a) the offsets corresponding to the onset front is determined, (b) the matching offset front members are found.

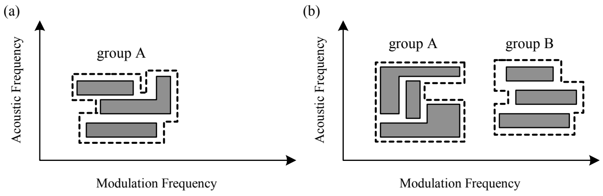

Figure 8 A graphical expression of the method of grouping segments in the modulation spectrogram domain; (a) for one speaker, (b) for two speakers.

3.3. Speech separation

In [17], a mask is presented for speech separation in the modulation spectrogram domain, assuming that the pitch ranges of the target and interference are known and that these ranges are the same in each subband. Our system extends this idea by allowing the value of the mask in each filter channel to depend on the estimated pitch range of that filter channel.

Consider a given signal $x(n)$ that is the sum of a target signal $x_{ts}(n)$ and an interference signal $x_{is}(n)$, sampled at f_s Hz, i.e., $x(n) = x_{ts}(n)+x_{is}(n)$. A proper mask should be estimated for segregating the target signal from the interference signal. In each filter channel k, the pitch ranges of the target and interfering speakers (obtained from the previous stage) are denoted by $PF_{ts}^k := [pf_{ts,\text{low}}^k, pf_{ts,\text{high}}^k]$ and $PF_{is}^k := [pf_{is,\text{low}}^k, pf_{is,\text{high}}^k]$, respectively. Also, $Q^k := \left\{ i \in \{0, \dots, I-1\} \quad \text{such that } (i.f_s)/(I.M) \in PF^k \right\}$ is defined as the set of modulation frequency indices of PF^k, i.e., a pitch range in the filter channel k.

To produce a frequency mask in each filter channel k, define the mean of the modulation spectral energy relative to a pitch range as the energy normalized by the wideness of that pitch range:

$$E^k = \left(\sum_{i \in Q^k} \left| X(k,i) \right|^2 \right) \Bigg/ (pf_{\text{high}}^k - pf_{\text{low}}^k) \tag{9}$$

The frequency mask is calculated, when the means of the modulation spectral energy of the target and interference speakers are compared in the following sense.

$$F^k = \frac{E_{ts}^k}{E_{ts}^k + E_{is}^k}, \tag{10}$$

Since there are artifacts associated with applying masks in the modulation frequency domain (see [22]), this domain is not preferable for modulation filtering in order to mask out the interference and reconstruct a time-domain signal. Instead, the frequency mask is transformed to the time domain. To this end, *a filter* *with linear phase* is constructed whose magnitude is F^k and the assigned linear phase is $\varphi^k(i) = i$. Then, the inverse DFT is taken

$$f^k(m) = \frac{1}{I}\sum_{i=0}^{I-1} F^k e^{j\phi^k(i)} e^{j2\pi mi/I}. \qquad (11)$$

The separated target signal is estimated by the convolution (over the variable m) of the obtained filter $f^k(m)$ with the modulator signal of the mixture signal $x(n)$ and then, multiplying by the carrier signal of the mixture signal

$$\tilde{X}(m,k) = \left[M(m,k) * f^k(m)\right] C(m,k), \qquad (12)$$

Finally, the separated target signal in the time domain is obtained by taking the inverse STFT of $\tilde{X}(m,k)$.

4. Evaluation

As mentioned earlier, our system estimates the pitch range and uses this range for the speech separation. In this section, we evaluate the proposed system in the processes of pitch range estimation and speech separation.

4.1. Pitch range estimation

First, the proposed system is evaluated in the pitch range estimation process with utterances chosen from the Lee's database [23] and a corpus of 100 mixtures of speech and interference [24], commonly used for CASA research, see, e.g., [13,25,26]. The corpus contains utterances from both male and female speakers. These utterances are mixed with a set of intrusions at different SNR levels. These intrusions are N0: 1 kHz pure tone; N1: white noise; N2: noise bursts; N3: cocktail party noise; N4: rock music; N5: siren, N6: trill telephone; N7: female speech; N8: male speech; and N9: female speech. These intrusions have a considerable variety; for example, N3 is noise-like, while N5 contains strong harmonic sounds. They form a realistic corpus for evaluating the capacity of a CASA system when it deals with various types of interference.

The signal $X(k, i)$ is the modulation spectrogram of an input signal that is digitized at a 16-kHz sampling rate. The parameters of the proposed system are set to M = 16 and K = 128. $w(n)$ is a Hanning window with length L = 64 (refer to Section 2). The STFT filter-bank has 128 filter channels, for which the center frequency of the kth filter channel is $\omega_k = 2\pi k/K$, k = 0,..., K-1.

Figure 9 shows the modulation spectrogram and the obtained segments for a typical speech frame, when the proposed system is applied. The speech signal is a mixture of target and interference with the overall SNR of 0 dB. We select a male speech, a white noise and a trill telephone as the interference. The results show that although the powers of the speech and interference signals are equal, the proposed method is still able to estimate the pitch range of the target speaker with a reasonable accuracy.

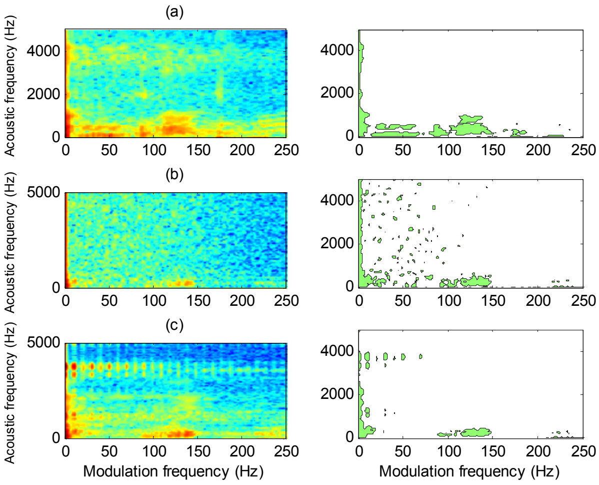

Figure 9 Modulation spectrogram and segments obtained for a mixture of (a) male speaker, (b) white noise, and (c) trill telephone. The input is shown in Figure 3a.

Figure 10 shows the average error percentage of the pitch range estimation by the proposed system on the above mixtures at different SNR levels. To determine the error percentage, we assign a two-element vector to the margins of each pitch range and find the root mean square error distance between the vectors corresponding to the true and estimated pitch ranges. As shown in Figure 10, the proposed system is able to estimate 79.9% of the target pitch range, even at -5 dB SNR. The estimation rate increases to about 96.1%, as the SNR increases to 15 dB.

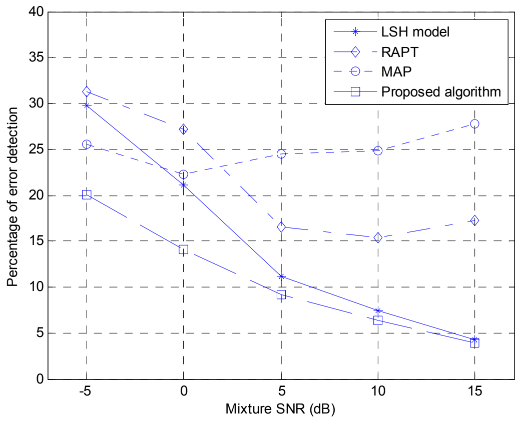

Figure 10 Percentage of pitch range estimation error for different SNR levels.

A reliable evaluation of the proposed system requires a reference range of the true pitch. However, such a reference is probably impossible to obtain from a noisy speech. We find the reference pitch range by framing the clean speech signal and calculating the pitch frequency in each frame.

The performance of the proposed method is compared with that of the Least Square Harmonic (LSH) technique [27], Robust Algorithm for Pitch Tracking (RAPT) [28], and the Maximum A Posterior (MAP) estimator [29]. RAPT and MAP are two standard pitch estimation algorithms. The LSH algorithm, derived in [27] for harmonic decomposition of a time-varying signal, estimates the harmonic amplitudes and phases, by solving a set of linear equations that minimizes the mean square error. The RAPT algorithm estimates the pitch frequency, by searching for local maxima in the autocorrelation function of the windowed speech signal and then, using a dynamic programming technique (see [28] for more details). The MAP approach [29] considers a harmonic model for the voiced speech so that each windowed signal is expressed with a generalized linear model whose basic functions depend on the fundamental frequency and number of harmonic partials.

Figure 10 also provides a comparison between the results of the pitch estimation using the mentioned four methods, in which the proposed system performs consistently better than the three standard methods, at all SNR levels. Although the performance of the LSH model (as the best performing one among the mentioned standard algorithms) is good at SNR levels above 10 dB, it drops quickly as SNR decreases, which shows that the proposed system is more robust to interference compared with the LSH model.

As mentioned in [29], MAP performs slightly better in low SNR's rather than high SNR's. In addition, RAPT fails to estimate the desired pitch period in low SNR's, because it mistakenly chooses sub-harmonic and harmonic partials instead of the true pitch period. The current scheme performs almost consistently in both high and low SNR's.

4.2. Voiced speech separation

A corpus of 100 mixtures composed of 10 target utterances mixed with 10 intrusions is recruited for assessing the performance of the system on voiced speech separation; these data are described in Section 4. 1. For comparison, the Hu and Wang system [14] and the spectral subtraction method [30] are employed. Performance of the voiced speech separation is evaluated using two measures commonly used for this propose [14]:

- The percentage of energy loss, P_{EL}, which measures the amount of the target speech excluded from the segregated speech.
- The percentage of residual noise, P_{NR}, which measures the amount of the intrusion included in the segregated speech.

P_{EL} and P_{NR} are error measures of a separation system, which are complementary indices for assessing the system performance. In addition, the SNR of the segregated voiced target (in dB) provides a good comparison between waveforms [14]:

$$SNR = 10\log_{10}\frac{\sum_n s^2(n)}{\sum_n\left[s(n)-\tilde{x}(n)\right]^2}, \tag{13}$$

where $\tilde{x}(n)$ is the estimated signal and $s(n)$ is the target signal before being mixed with the intrusion.

The results of our system are shown in Figure 11. Each point in the figures represents the average value of 100 mixtures in the complete test corpus at a particular SNR level. Figure 11a, b shows the percentage of energy loss and noise residue. Since the goal here is to segregate the voiced target, the P_{EL} values are only defined for the target energy at the voiced frames of the target.

As shown in Figure 11, the proposed system segregates 78.9% of the voiced target energy at -5 dB SNR and 99% at 15 dB SNR. At the same time, at -5 dB, 15.9% of the segregated energy belongs to intrusion. This number drops to 0.7% at 15 dB SNR. Figure 11c shows the SNR of the segregated target. This system obtains an average 7.5 dB gain in SNR when the mixture SNR is -5 dB. This gain increases to 14.3 dB, when the mixture SNR is 15 dB. As shown in the figure, the segregated target loses more target energy (Figure 11a), but contains less interference as well (Figure 11b).

Figure 11 also shows the performance of the system proposed by Hu and Wang for voiced speech separation

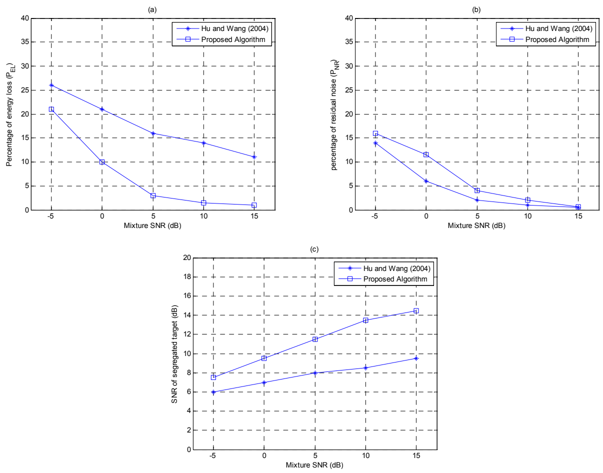

Figure 11 Results of voiced speech separation. **(a)** Percentage of energy loss on voiced target. **(b)** Percentage of noise residue. **(c)** SNR of segregated voiced target.

[14], which is a representative of CASA systems. As shown in the figure, the Hu and Wang's system yields a lower percentage of noise residues (Figure 11b), but has a much higher percentage of target energy loss (Figure 11a, c). Nevertheless, it should be noted that our system significantly improved the P_{EL} (in Figure 11a, see, e.g., by around 11 and 10% improvement at 0 and 15 dB, respectively), which leads to much less signal distortion. The price paid for this is a slightly increase in P_{NR}, as depicted in Figure 11b (e.g., by around 6 and 0.5% increase at 0 and 15 dB, respectively).

The average SNR for each intrusion is shown for the proposed system in Figure 12 in comparison with that of the original mixtures, Hu and Wang's system, and a Spectral Subtraction Method, which is a standard method for speech enhancement [30] (see also [14]). The proposed system performs consistently better than Hu and Wang's system and spectral subtraction. In average, the proposed system obtains a 16.85 dB SNR gain, which is about 1.92 dB better than Hu and Wang's system and 8.4 dB better than the Spectral Subtraction.

To help the reader recognize the real difference in the performance, a file is prepared including sample audio mixture signals (target speech signal + interference signal) and the results of the separation using the Spectral Subtraction, Hu and Wang, and the proposed systems. The file is available at http://ee.yazduni.ac.ir/sprl/ASP-AM-SampleWaves.ppt.

5. Discussions and conclusions

One of the major challenges in speech enhancement is the separation of a target speech from an interference signal of the same type. The accuracy of the CASA methods in single channel speech separation depends on the correctness of the pitch frequency estimation of two simultaneous speakers because the proper mask in the T-F domain for the speech separation is produced in association with the estimated pitch frequency.

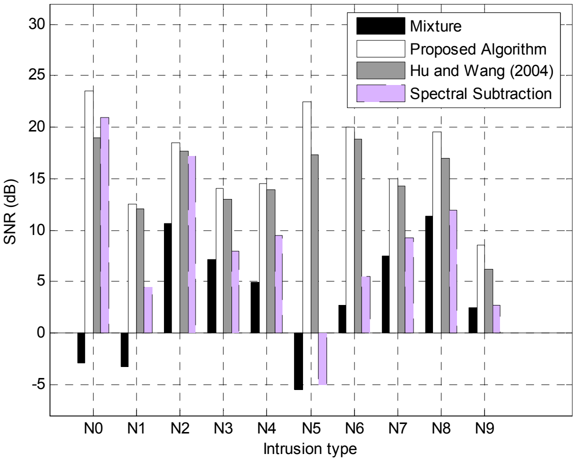

Figure 12 SNR results for segregated speech and original mixtures for a corpus of voiced speech and various intrusions.

In this article, a single channel speech separation system is proposed that estimates the pitch range of one or two speakers and segregates the target speech from the interference. The pitch range is estimated using the onset and offset algorithm considering the distribution of speaker energy in the modulation spectrogram domain. When the target and interference speakers are either male or female, the methods for pitch frequency estimation encounter large errors because of close pitch frequency values. Therefore, CASA methods that employ the pitch frequency as their main feature for speech separation face difficulties. In contrast, a main novelty of the present algorithm is the estimation of pitch range based on short time-frames of the mixture signal. The constructed mask for speech separation depends on the pitch range estimated independently in each subband. As shown by the evaluation results, major portions of the voiced target speech are separated from the interfering speech using this mask. In addition, the proposed system can separate the unvoiced portions that are quasi-periodic because of the proximity of voiced portions.

The proposed algorithm is robust to interference and produces good estimates of both pitch range and voiced speech, even in the presence of strong interference. Systematic evaluation shows that the proposed algorithm performs significantly better than the mentioned CASA and speech enhancement systems.

Silent gaps and other interference-masked intervals are usually included in natural speech utterances. In practice, the utterance across such time-intervals should be grouped. This is a sequential grouping problem [5,6] whose segments or masks can be obtained using the speech recognition in a top-down manner (also, limited to non-speech interference) [11] or the speaker recognition trained by speaker models [31]. However, the proposed algorithm does not encounter this problem of sequential grouping because it operates in the modulation spectrogram domain.

In terms of computational complexity, the main cost of the proposed algorithm arises from determining segments in modulation spectrogram for pitch range estimation. The estimation of the mask and convolution for speech separation consumes a small fraction of the overall cost. Both tasks (pitch range estimation and speech separation) are implemented in the frequency domain, so the computational complexity is $O(N\log N)$, where N is the number of samples in the input signal. These operations should separately be performed for each subband. On the other hand, since feature extraction takes place independently in different subbands, substantial speedup can be achieved through parallel computing.

For future work, the proposed algorithm can be improved by iterative estimation of pitch range and speech separation. The algorithm can include a specific method to jump-start the iterative process, which gives an initial estimate of both pitch range and mask with reasonable quality. In general, the performance of the algorithm depends on the initial estimate of pitch range; better initial estimates would lead to better performance. Even with a poor estimate of pitch range, which is unavoidable in very low SNR conditions, the proposed algorithm improves the initial estimate during the iterative process.

Author details

[1]Speech Processing Research Lab (SPRL), Electrical and Computer Engineering Department, Yazd University, Yazd, Iran [2]Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran [3]Image Analysis Laboratory, Department of Radiology, Henry Ford Health System, Detroit, MI, USA [4]Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran

Competing interests

The authors declare that they have no competing interests.

Received: 7 May 2011 Accepted: 17 March 2012
Published: 17 March 2012

References

1. RP Lippmann, Speech recognition by machines and humans. Speech Commun. **22**, 1–16 (1997). doi:10.1016/S0167-6393(97)00021-6
2. JJ Sroka, LD Braida, Human and machine consonant recognition. Speech Commun. **45**, 410–423 (2005)
3. A de Cheveigne, in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, ed. by Brown GJ, Wang DL (Wiley & IEEE, Hoboken, NJ, 2006), pp. 45–79
4. S Dubnov, J Tabrikian, M Arnon-Targan, Speech source separation in convolutive environments using space-time-frequency analysis. EURASIP J Appl Signal Process Article 38412. **2006**, 11 (2006)
5. AS Bregman, *Auditory Scene Analysis* (MIT, Cambridge, MA, 1990)
6. Brown GJ, Wang DL (eds.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley & IEEE, Hoboken, NJ, 2006)
7. M Buchler, S Allegro, S Launer, N Dillier, Sound classification in hearing aids inspired by auditory scene analysis. EURASIP J Appl Signal Process. **18**, 2991–3002 (2005)
8. G Hu, D Wang, A Tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans Audio Speech Lang Process. **18**(8), 2067–2079 (2007)
9. Y Shao, S Srinivasan, Z Jin, D Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition. Comput Speech Lang. **24**, 77–93 (2010). doi:10.1016/j.csl.2008.03.004
10. MH Radfar, RM Dansereau, A Sayadiyan, A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation. EURASIP J Audio Speech Music Process Article 84186, **2007**, 15 (2007)
11. J Barker, M Cooke, D Ellis, Decoding speech in the presence of other sources. Speech Commun. **45**, 5–25 (2005). doi:10.1016/j.specom.2004.05.002
12. Y Shao, DL Wang, Model-based sequential organization in cochannel speech. IEEE Trans Acoust Speech Signal Process. **14**, 289–298 (2005)
13. GJ Brown, M Cooke, Computational auditory scene analysis. Comput Speech Lang. **8**, 297–336 (1994). doi:10.1006/csla.1994.1016
14. G Hu, DL Wang, Monaural speech separation based on pitch tracking and amplitude modulation. IEEE Trans Neural Net. **15**, 1135–1150 (2004). doi:10.1109/TNN.2004.832812
15. M Wu, DL Wang, GJ Brown, A multipitch tracking algorithm for noisy speech. IEEE Trans Speech Audio Process. **11**, 229–241 (2003). doi:10.1109/TSA.2003.811539
16. J Le Roux, H Kameoka, N Ono, A de Cheveigne, S Sagayama, Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments. IEEE Trans Audio Speech Lang Process. **15**, 1135–1145 (2007)
17. SM Schimmel, LE Atlas, K Nie, Feasibility of single channel speaker separation based on modulation frequency analysis, in *Proc IEEE International Conference on Acoustics, Speech and Signal Processing, Hawaii, USA*. **4**, 605–608 (2007)
18. SM Schimmel, (Dissertation, University of Washington, 2007)
19. L Atlas, SA Shamma, Joint acoustic and modulation frequency. EURASIP J Appl Signal Process. **2003**(7), 668–675 (2003). doi:10.1155/S1110865703305013
20. G Hu, DL Wang, Auditory segmentation based on onset and offset analysis. IEEE Trans Audio Speech Lang Process. **15**(2), 396–405 (2007)
21. R Drullman, JM Festen, R Plomp, Effect of temporal envelope smearing on speech reception. J Acoust Soc Am. **95**, 1053–1064 (1994). doi:10.1121/1.408467
22. SM Schimmel, LE Atlas, Coherent envelope detection for modulation filtering of speech, in *Proc IEEE International Conference on Acoustics, Speech and Signal Processing, Pennsylvania, USA*, 221–224 (2005)
23. TW Lee, Blind source separation: audio examples (1998). http://www.snl.salk.edu/~tewon/Blind/blind_audio.html. Accessed 4 May 2011
24. MP Cooke, *Modeling Auditory Processing and Organization* (Cambridge University Press, Cambridge, 1993)
25. LA Drake, (Dissertation, University of Northwestern, 2001)
26. DL Wang, GJ Brown, Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans Neural Netw. **10**, 684–697 (1999). doi:10.1109/72.761727
27. Q Li, L Atlas, Time-variant least-squares harmonic modeling, in *Proc IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong*. **2**, 41–44 (2003)
28. D Talkin, A robust algorithm for pitch tracking (RAPT), in *Speech Coding and Synthesis*, ed. by Paliwal KK, Klein WB (Elsevier, NewYork, NY, 1995), pp. 495–518
29. J Tabrikian, S Dubnov, Y Dickalov, Maximum a posterior probability pitch tracking in noisy environments using harmonic model. IEEE Trans Speech Audio Process. **12**, 76–87 (2004). doi:10.1109/TSA.2003.819950
30. X Huang, A Acero, HW Hon, *Spoken Language Processing: A Guide to Theory, Algorithms, and System Development* (Prentice Hall PTR, Upper Saddle River, NJ, 2001)
31. Y Shao, (Dissertation, University of Ohio State, 2007)

doi:10.1186/1687-6180-2012-67
Cite this article as: Mahmoodzadeh *et al.*: Single channel speech separation in modulation frequency domain based on a novel pitch range estimation method. *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:67.