Monaural Voiced Speech Segregation Based on Pitch and Comb Filter

Xueliang Zhang\(^1\) and Wenju Liu\(^2\)

\(^1\) Compute Science Department, Inner Mongolia University, Huhhot, China, 010021
\(^2\) National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences Beijing, China, 10019

\texttt{cszxl@imu.edu.cn}
\texttt{lwj@nlpr.ia.ac.cn}

\textbf{ABSTRACT}

The correlogram is an important mid-level representation for periodic sounds which is widely used in sound source separation and pitch detection. However, it is very time consuming. In this paper, we presented a novel scheme for monaural voiced speech separation without computing correlograms. The noisy speech is firstly decomposing into time-frequency units. Pitch contour of the target speech is extracted according to the zero crossing rate of the units. Then we applied a comb filter to label each unit as target speech or intrusion. Compared with previous correlogram-based method, the proposed algorithm saves computing time and also yields better performance.

\textbf{Index Terms—} Sound separation, Computational auditory scene analysis, Correlogram

\section{1. INTRODUCTION}

Human auditory system has superior capacity to focus on a single talker among a mixture of conversations and background noises. By exploring the process of human to sound perception, Bregman proposed a theory called \textit{Auditory Scene Analysis} (ASA) \cite{Bregman}. ASA inspired research in the domain referred to as Computational Auditory Scene Analysis (CASA) \cite{CASA}.

The general framework of CASA-based separation systems has two main stages: segmentation and grouping. In segmentation, the acoustic input is decomposed into sensory segments, each of which should originate from a single source. In grouping, those segments that likely come from the same source are grouped together according to the grouping cues. For voiced speech separation, pitch or fundamental frequency (F0) is one of the most important cues. Given the F0, systems can utilize harmonicity principle to group the segments in different frequency regions. Harmonicity principle shows that F0 and its overtones are perceived as single source by human beings.

A well-established representation for harmonic structure is a correlogram \cite{correlogram}, which has been adopted by many CASA systems. Input signal is firstly decomposed into multiple channels by auditory filterbank. The correlogram is a running autocorrelation of the signal within a certain period of time in each filter channel. The periodicity of the signal is represented by the corresponding autocorrelation function (ACF). Separation systems \cite{previous} also employed cross channel correlation of correlogram for segmentation.

Although CASA is faced with many difficulties (such as sequential organization, unvoice segregation), it can be still used in some practical applications (e.g. extracting singing voice from musical sound and musical instruments separation given the ground truth F0). The truth F0 can be obtained by user’s singing or MIDI files. However, computing correlogram is very time consuming which limits CASA for these kinds of applications.

In this paper, we proposed a novel scheme to separate the voiced speech from intrusions in monaural situation instead of computing correlograms. At first, input signal is decomposed into time-frequency units. Then, the units are merged into several segments. Pitch of the target is extracted by segment-based method. After that, each unit is labeled as target or intrusion according to harmonicity principle and amplitude modulation (AM) criterion \cite{Bregman} which will be introduced later. Finally, the labeled units are separated into foreground or background. Being different with previous systems, critical parts (segmentation, pitch estimation and unit labeling) of the proposed system aren’t based on correlograms and run much faster.

The rest of paper is organized as follows. In section 2, we describe the details of each stage. In section 3, time complexity of the algorithms is analyzed. Signal to noise ratio (SNR) of separated speech and computing time are reported in section 4. A conclusion is given in section 5.

\section{2. SYSTEM DESCRIPTION}

The proposed system has four stages as shown in Figure 1.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{system_diagram.png}
\caption{Systematic diagram}
\end{figure}

\subsection{2.1. Front-End processing}

In the typical Front-End processing (e.g. in \cite{previous}), input signal
\( x(t) \) is decomposed by a gammatone filterbank with 128 channels whose center frequencies are quasi logarithmically spaced from 80 Hz to 5 kHz and bandwidths equal to the equivalent rectangle bandwidth (ERB). Instead, the proposed system decomposes the signal by forward-backward filtering. The reason will be given later. Specifically, \( x(t) \) is first passed through the gammatone filterbank. The outputs are time reversed and re-filtered by the gammatone filterbank. Then the filter outputs are time reversed again. After that, phase delay of output is compensated in each channel. Because forward-backward filtering causes the actual bandwidth narrower than standard gammatone filter, we enlarge the bandwidths to 1.6 times of the ERB.

AM of gammatone filter output is extracted by band-pass filtering the Hilbert envelope which is a conventional method \[2\]. Considering the plausible pitch range of speech, the bandpass is set from 50 Hz to 550 Hz. Gammatone filter method \[2\]. Considering the plausible pitch range of speech, we enlarge the bandwidths to 1.6 times of the ERB.

The output of each channel is then divided into 20 ms time frames and 10 ms time shift. In the following part, \( u_{\text{norm}} \) is to denote a time-frequency (T-F) unit for frequency channel \( c \) and time frame \( m \). As in the Hu and Wang model \[4\], we employ different methods to segregate resolved and unresolved units. The resolved one is defined as being dominated by single harmonic and unresolved one is defined as being dominated by several harmonics. In order to discriminate the resolved and unresolved units, we introduce a feature called carrier to envelope energy ratio (CER) which is calculated as (1). \( u_{\text{norm}} \) is termed as resolved if \( R_{\text{avg}}(c,m)>\theta_c \); otherwise it is termed as unresolved. Motivation is that when unit is dominated by several harmonics, the AM is relative strong and it leads to a small value for \( R_{\text{avg}} \).

\[
R_{\text{avg}}(c,m) = \log \left( \frac{\sum_{n=0}^{W} g(c,mT + n)^2}{\sum_{n=0}^{W} e(c,mT + n)^2} \right)
\]  

where, \( T = 160 \) and \( W = 320 \) corresponding to 10 ms time shift and 20 ms time frame; the frequency sampling rate is 16 kHz.

Segmentation plays an important role in CASA system. Each segment consists of spatially (time and frequency) continuous units which are generated according to cross-channel correlation. In stead of computing it by corrologram (as in \[4\]\[5\]), the cross channel correlation is computed directly on gammatone filter outputs by (2) and it is the reason that we use forward-backward filtering to compensate the phase delay.

\[
C_g(c,m) = \sum_{n=0}^{N-1} g(c,mT + n) \times \hat{g}(c+1,mT + n)
\]  

where, \( \hat{g}(c,t) \) is zero-mean and unity-variance version of filter response at channel \( c \) and window \( m \).

In addition, we compute the zero crossing rate of gammatone filter output with positive slope in each unit, termed as \( Z(c,m) \) for \( u_{\text{norm}} \). ZCR is used for pitch detection in the following section.

2.2. Pitch estimation

In this subsection, the continuous pitch contour is extracted based on segments and ZCR of units. We know that ZCR does not work well for complex waveform. Therefore, the units used in pitch estimation should be dominated by signal harmonic. The resolved units tend to meet this requirement. To improve the accuracy, the resolved units are further selected by the segmentation. Specifically, the units are firstly selected as candidates when CER exceeds a threshold \( \theta_c = 1.82 \) and cross channel correlation is larger than 0.98. Then the neighboring candidates are merged into segments. The segments shorter than 30 ms are removed since which unlikely arise from target speech. The remaining units in longer segments are used for pitch estimation.

We employ cosine function as substitution for pitch detection whose frequency is set to the ZCR. In the selected units, cosine function has similar shape with autocorrelation function. The rest of pitch detection is similar with the Hu and Wang model. The dominant pitch at each frame is shown by the maximum peak of summary cosine functions. Then, we use the longest segment and dominant pitch as a criterion to segregate each segment into foreground and background (details can be found in \[4\]). Different from the process in \[4\], pitch estimation is based on the longest segment in foreground and its harmonic order. Harmonic order shows that the segment is dominated by which harmonic. It is given by (3).

\[
H = \arg \max_m \left\{ \sum_{m \in \text{long}} SCF(m) \right\}
\]  

where, \( S_{\text{long}} \) stands for the longest segment in foreground; \( Z(c,m) \) is the zero crossing rate at unit \( u_{\text{norm}} \); \( SCF(m) \) is summary cosine function of units in foreground at frame \( m \).

The pitch period at frame \( m \) is determined by (4)

\[
P(m) = \arg \max \left\{ SCF^1(m,\tau) \right\}
\]  

where, \( \tau \in [2\ ms, 12.5\ ms] \) corresponding to the range \( [80\ Hz, 500\ Hz] \); \( SCF^1(m,\tau) \) is summary cosine functions with the range \( [H\pi-\pi/2, H\pi+\pi/2] \) of units in longest segments.

2.3. Unit labeling

We use estimated pitch period to label the units as “target” or “intrusion”. The labeled units will be segregated into foreground and background in next subsection.

As in the Hu and Wang model, resolved and unresolved T-F units are treated differently. For the resolved one, it is
labeled according to *harmonicity* principle. If response frequency is multiple of the estimated pitch, the resolved T-F unit is labeled as target. For the unresolved one, it is labeled according to *AM* criterion. AM criterion pointed out that if a filter responding to multiple harmonics of a single harmonic sound source, the response envelope fluctuates at the rate of $F_0$ of the source. Therefore, the unresolved T-F unit is labeled as target if the AM rate equals to the estimated pitch.

To measure these two criterions, we employ an IIR comb filter with sieves at $F_0$ and its overtones to filter the flattened output of gammatone filter $g_0(c,t)$ or its flattened envelope $e_0(c,t)$. The flattened signals are obtained by (5). This process can be viewed as a simplified simulation for the automatic gain control which is one of the functions of cochlear.

$$ r_g(c,t) = \text{sgn}[r(c,t)] \times |r(c,t)|^\rho $$  \hspace{1cm} (5)

where, $r(c,t)$ stands for output of gammatone filter or its envelope at channel $c$; $r_0(c,t)$ is the flattened signal; $\rho$ is the compression rate, here $\rho=0.1$.

Then the flattened signals are passed through the comb filter

$$ r_f(c,t) = r_g(c,t) + \alpha \times r_g(c,t-P(t)) $$  \hspace{1cm} (6)

where, $P(t)$ is the pitch period at time $t$ which is obtained by linear interpolation of the frame pitch period; $r_g(c,t)$ stands for comb filter output; $\alpha=0.55$.

Unit $u_{cm}$ is labeled as target if most of the flattened signals passing through the comb filter, i.e. if the relative energy of comb filter output is above a certain threshold $\theta_t$; otherwise $u_{cm}$ is labeled as intrusion.

$$ \log \left[ \frac{\sum_{n=-\infty}^{\infty} r_{c,mT+n}^2}{\sum_{n=-\infty}^{\infty} r_{c,mT+n}} \right] > \theta_t $$  \hspace{1cm} (7)

For resolved T-F unit, $r(c,t)$ in (5) stands for the output of gammatone filter $g_0(c,t)$. Similarly, $r_0(c,t)$ and $r_g(c,t)$ stands for $g_0(c,t)$ and $g_0(c,t)$ respectively. For unresolved T-F unit, $r(c,t)$ in (5) stands for the envelope of gammatone filter output $e_0(c,t)$. The threshold $\theta_t=0.55$ for resolved units and $\theta_t=0.45$ for unresolved units.

### 2.4. Separation and synthesis

To directly use labeling information as the final decision will lead to some errors. Hence, previous method [4] provided a separation method based on segmentation. Here, we take the similar process with different details.

**a)** The resolved T-F unit separation is based on the segments generated in Section 2.2. These segments are firstly marked as matched or mismatched on each frame. Specifically, if more than 50% units on a frame are labeled as target, we call the segment matched on this frame. If more than half of the segment’s frames marked as matched, it is grouped into foreground; otherwise it is grouped into background. In foreground, the units labelled as intrusion are merged into new segments. Then the segments larger than 30 ms are moved into background.

**b)** For unresolved T-F units, they respond to several frequency components. If dominated by target voiced speech, their AM rate equals to the $F_0$. Therefore, the flattened envelope is used in (5) and (6). However, the large value in (6) doesn’t mean AM rate equals to $F_0$. The units are possibly dominated by noise with fractional $F_0$ (e.g. $F_0/2$, $F_0/3$...). To eliminate the errors, we compute the ZCR of unresolved T-F units on the comb filter output. The segments consisting of unresolved T-F units are formed by the spatially continuous candidates with distance between ZCR and $F_0$ less than 50%. The segments longer than 30 ms are grouped into foreground.

**c)** The units labelled as target, not in foreground, are merged iteratively into its adjacent segments in foreground. The rest are grouped into background.

Finally, the units in foreground are used to synthesize the waveform of separated target speeches.

### 3. ANALYSIS OF RUN-TIME COMPLEXITY

In this section, we analyze the run-time complexity of the proposed algorithm and compare it with the Hu and Wang model. As a typical correlogram-based speech separation system, the Hu and Wang model [4] has much better performance than previous systems. The innovations of the Hu and Wang model are that: 1) different separation method for resolved and unresolved harmonics; 2) separation based on segmentation; 3) pitch detection in noisy environment. Correlogram plays a vital role in each stage.

Because the entire separation systems are relative complicated, we only compared the major processes in each stages. From table 1, it can be seen that computing correlograms is the bottleneck of the Hu and Wang system. To accelerate the computation, autocorrelation could be done in frequency domain [2]. Then the complexity of computing correlograms is $O(CL\log W)$, where $W$ is frequency domain and its complexity is $O(CL)$. In our algorithm, the complexity of those two counterparts is $O(CL)$. Specific running time of the algorithms is shown in next section.

### 4. SYSTEM EVALUATION

The proposed scheme is evaluated on a corpus of 100 mixtures composed of ten voiced utterances mixed with ten different kinds of intrusions collected by Cooke [6] which is widely used to evaluate the separation systems. In the dataset, ten voiced utterances have continuous pitch nearly
throughout whole duration. And the intrusions are ten different kinds of sounds including 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. Ten voiced utterances are regarded as targets. Frequency sampling rate of the corpus is 16 kHz.

As commonly used objective performance measure for separation systems [4][5], signal to noise ration (SNR) is chosen. Its computation is as follows:

\[
SNR = 10 \log_{10} \left( \frac{\sum R(t)^2}{\sum [R(t) - S(t)]^2} \right)
\]

(8)

where, \(R(t)\) is the clean speech and \(S(t)\) is the synthesized waveform by segregation systems.

In table 2, each value represents the average SNR for one intrusion mixed with ten target speeches. The average of overall intrusions is shown in last row. As shown in table, our algorithm improves SNR for most of the intrusions and produces a gain of 0.7 dB over the Hu and Wang model.

For further comparison, we replace estimated pitch by true pitch in the Hu and Wang model (termed as TP-HW) and the proposed algorithm (termed as TP-Pro). The true pitch is obtained by performing the algorithm on clean speech. The performances are listed in table 2. It can be seen that TP-HW produces a gain of 0.35 dB compared with the original Hu and Wang model. While the TP-Pro produces a gain of 0.17 dB. Although we didn’t compare the pitch estimation of both algorithms, the proposed pitch estimation is at least no worse than the method in the Hu and Wang model. The SNR gap between TP-HW and TP-Pro is 0.53 dB.

To compare the computing time, both of the proposed algorithm and the Hu and Wang model are implemented by C language and run on the PC platform with 1.6 GHz CPU and 3 GB storage memory. The implementation of the Hu and Wang model is provided by Prof. Deliang Wang. We also accelerate the Hu and Wang model by computing correlograms and bandpass filtering in frequency domain which is termed as accHW. Results of computing time are listed in table 3.

From table 3, we can see that total duration of 100 mixtures is 168.3 seconds. And the computing time of the Hu and Wang is 14.6 times of real time. For the accelerated Hu and Wang model, the computing time is 6.33 times of real time. It saves 57% computing time. While, the total computing time of the proposed system is 2.23 times of real time. Compared with original Hu and Wang model and accelerated Hu and Wang model, the proposed method saves 84.8% and 64.8% of computing time respectively.

5. DISCUSSION AND CONCLUSION

In this paper, we propose a novel algorithm for monaural voiced speech separation which avoids computing the correlograms. Segmentation, pitch detection and segregation are implemented in an efficient way by the novel scheme. Compared with the typical correlogram-based algorithm Hu and Wang model, the proposed scheme achieves better performance and saves computing time.

6. REFERENCE