INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

A Signal Processing Approach for Speaker Separation using SFF Analysis

Nivedita Chennupati[1], *BHVS Narayana Murthy*[1,2] *and B. Yegnanarayana*[1]

[1]Speech Processing Lab, International Institute of Information Technology, Hyderabad, India.
[2]Research Centre Imarat, Hyderabad, India.

nivedita.chennupati@research.iiit.ac.in, bhvsnm@rcilab.in, yegna@iiit.ac.in

Abstract

Multi-speaker separation is necessary to increase intelligibility of speech signals or to improve accuracy of speech recognition systems. Ideal binary mask (IBM) has set a gold standard for speech separation by suppressing the undesired speakers and also by increasing intelligibility of the desired speech. In this work, single frequency filtering (SFF) analysis is used to estimate the mask closer to IBM for speaker separation. The SFF analysis gives good temporal resolution for extracting features such as glottal closure instants (GCIs), and high spectral resolution for resolving harmonics. The temporal resolution in SFF gives impulse locations, which are used to calculate the time delay. The delay compensation between two microphone signals reinforces the impulses corresponding to one of the speakers. The spectral resolution of the SFF is exploited to estimate the masks using the SFF magnitude spectra on the enhanced impulse-like sequence corresponding to one of the speakers. The estimated mask is used to refine the SFF magnitude. The refined SFF magnitude along with the phase of the mixed microphone signal is used to obtain speaker separation. Performance of proposed algorithm is demonstrated using multi-speaker data collected in a real room environment.

Index Terms: Multi-speaker separation, single frequency filtering (SFF), time delay estimation, binary mask.

1. Introduction

Multi-speaker separation is useful in different group-gathering scenarios such as cock-tail party and meetings. The goal of source separation can be redefined as estimating a mask nearer to the ideal binary mask (IBM) from the mixed speech signals, where the target and secondary source signals are not available separately [1]. The mask estimation is generally attempted using the mixed speech data from one or more spatially distributed microphones. Single channel separation algorithms are computationally intensive, and depend on the robustness of pitch tracking algorithms [2]. Recently, people have shown good separation using machine learning techniques (such as DNNs) [3]. Such techniques use large amount of training data and supervised learning.

In this paper, a signal processing algorithm that exploits the high SNR regions of speech corresponding to a speaker is used for estimation of the mask. The high SNR regions in the time (corresponding to GCIs) [4] and in the frequency (corresponding to harmonics) [5] are estimated using single frequency filtering (SFF) analysis [6]. The SFF analysis and synthesis procedures are described briefly in Section 2. The algorithm for speaker separation is presented in Section 3.

2. SFF Analysis and Synthesis

In the single frequency filtering (SFF) analysis, the speech signal ($x[n]$) is shifted in frequency depending on the choice of frequency (f_k) by multiplying with a complex exponential $e^{-j\frac{2\pi\bar{f}_k}{f_s}n}$, where $\bar{f}_k = \frac{f_s}{2} - f_k$. The resulting signal ($\tilde{x}[k,n]$) is filtered by a near ideal resonator ($\frac{1}{1+0.995z^{-1}}$) at the highest frequency, namely, $f_s/2$, where f_s is the sampling frequency. The output of the filter is given by

$$y[k,n] = -ry[k,n-1] + \tilde{x}[k,n]. \tag{1}$$

$y[k,n]$ is extracted for different values of f_k, where $k = 0,1,2,\ldots K-1$, and K is the total number of filters covering the range 0 to f_s using a spacing of 5 Hz. Let

$$y[k,n] = v[k,n]e^{j\phi[k,n]}, \tag{2}$$

The SFF envelope ($v[k,n]$) and phase ($\phi[k,n]$) of the filtered output at the k^{th} frequency are given by

$$v[k,n] = \sqrt{y_r^2[k,n] + y_i^2[k,n]}, \tag{3}$$

$$\phi[k,n] = \tan^{-1}\left(\frac{y_i[k,n]}{y_r[k,n]}\right), \tag{4}$$

where $y_r[k,n]$ and $y_i[k,n]$ are the real and imaginary parts of $y[k,n]$, respectively. The $v[k,n]$ at each instant of time gives SFF magnitude spectrum.

The speech signal can be reconstructed by shifting the outputs at $f_s/2$ to the corresponding frequencies and summing them. The reconstructed signal is given by

$$s_r[n] = \frac{1}{K}\sum_{k=0}^{K-1} v[k,n]e^{j\phi[k,n]}e^{j\frac{2\pi\bar{f}_k}{f_s}n}. \tag{5}$$

It is observed that the mean square error between the original and reconstructed signals is negligible [7].

3. Speaker Separation Algorithm

Fig. 1 shows the block diagram of the proposed algorithm for binary mask estimation consisting of two stages. In Stage 1, the excitation signals of the individual speakers are estimated from the two channel data. The excitation signal of voiced regions in speech signal is considered as a sequence of impulses with period corresponding to pitch of the speaker. The impulse-like excitations appear as changes in the energy for short intervals of time. This can be observed in the gain contours derived from the SFF envelopes [4]. The gain of the SFF envelopes is given by,

$$\mu[n] = \frac{1}{K}\sum_{k=0}^{K-1} v[k,n]. \tag{6}$$

To highlight the impulse-like excitation, the first order difference is obtained on the gain contours. The difference of the gain contour from multi-speaker data will have impulses corresponding to all the speakers in the signal. To emphasize the

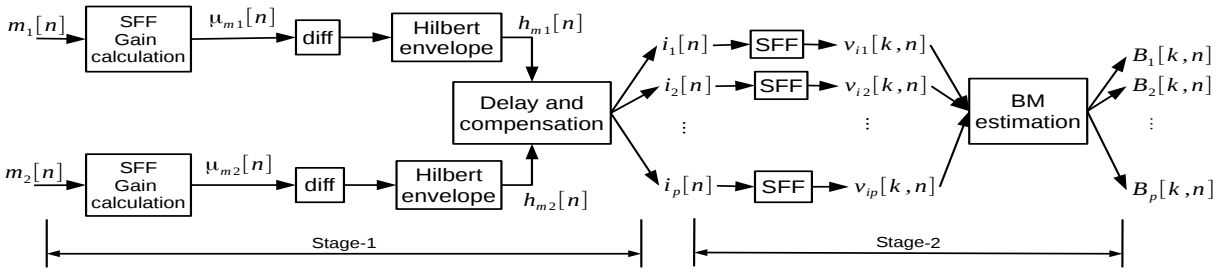

Figure 1: *Block diagram for binary mask estimation.*

impulses, Hilbert envelope (HE) is calculated on the difference of the gain contours. The HEs derived for the two microphones are used for calculating the time delays and the binary masks.

The segmental cross-correlation function is obtained using the derived HEs for a segment size of 100 ms and a shift of 5 ms. The maximum delay in the cross-correlation function is set to 6 ms which is dependent on the maximum separation between the microphones. The peak location in the cross-correlation function gives the time delay for a segment. Histogram is plotted on the estimated segmental delays. The number of prominent peaks in the histogram gives the number of speakers, and their locations give the time delay. If the delay is negative, it means that the signal reached the microphone 1 earlier than the microphone 2, and vice-versa for positive delay.

Once the delay corresponding to a speaker is computed and compensated, the impulse locations corresponding to the direct component of the speaker are coherent in the two microphone signals and the impulses due to reflections and other speakers are incoherent. To enhance the impulses corresponding to a speaker, the HEs derived from the two microphone signals are multiplied after compensating for the delays, and a square root is taken on the result.

In Stage 2, the SFF analysis is done on the enhanced sequence of impulses corresponding to each speaker. The computed SFF magnitude spectra highlights the harmonics of a particular speaker in the voiced regions. Binary mask (BM) is computed by comparing the calculated SFF magnitude spectra obtained on the enhanced impulse-like sequence. For a three speaker case, the BM corresponding to one of the speakers can be estimated as follows:

$$B_1[k,n] = B_{12}[k,n] * B_{13}[k,n], \tag{7}$$

where

$$B_{12}[k,n] = \begin{cases} 1, & if\ v_{i1}[k,n] \geq v_{i2}[k,n] \\ 0, & otherwise \end{cases} \tag{8}$$

$$B_{13}[k,n] = \begin{cases} 1, & if\ v_{i1}[k,n] \geq v_{i3}[k,n] \\ 0, & otherwise \end{cases} \tag{9}$$

and $v_{i1}[k,n]$, $v_{i2}[k,n]$ and $v_{i3}[k,n]$ are the SFF magnitude spectra derived from the enhanced impulse sequences corresponding to the three speakers. For a two speaker case, the BM is $B_{12}[k,n]$. Similarly, masks can be calculated for other speakers in a mixture.

A speaker's signal can be reconstructed from the mixed signal using the following steps.

- Assume that speaker 1 is nearer to microphone 1. The SFF analysis on microphone 1 signal gives magnitude ($v_{m1}[k,n]$) and phase ($\phi_{m1}[k,n]$) of the mixed signal.
- The mixed signal magnitude is modified by multiplying with the estimated binary mask ($B_1[k,n]$) corresponding to speaker 1.

$$\hat{v}_{m1}[k,n] = v_{m1}[k,n] * B_1[k,n]. \tag{10}$$

- Enhanced signal ($s_{1r}[n]$) corresponding to speaker 1 is reconstructed using SFF synthesis procedure by using the mixed phase ($\phi_{m1}[k,n]$) and modified magnitude ($\hat{v}_{m1}[k,n]$).
- SFF is applied on the enhanced signal to get modified phase corresponding to speaker 1.
- The modified phase and modified magnitude are used to reconstruct the enhanced signal after iteration.

4. Results

The results show good separation when competing speakers are equally loud. In some cases, the dominant speaker is separated well compared to the other speakers. Some examples of multi-speaker separation are available at `https://researchweb.iiit.ac.in/~nivedita.chennupati/multispksep/`.

5. Acknowledgements

The first author would like to thank Tata Consultancy Services(TCS) for supporting her PhD work.

6. References

[1] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, 2005, pp. 181–197.

[2] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067 – 2079, Nov. 2010.

[3] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, May 2017.

[4] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52 – 63, Feb. 2017.

[5] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *INTERSPEECH*, 2016, pp. 2155–2159.

[6] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, April 2015.

[7] N. Chennupati, S. R. Kadiri, and B. Yegnanarayana, "Intelligibility improvement of speech in noise using single frequency filtering approach," *submitted to JASA*, 2017.