Interspeech 2018
2-6 September 2018, Hyderabad

Determining Speaker Location from Speech in a Practical Environment

BHVS Narayana Murthy[1], *J V Satyanarayana*[2], *B. Yegnanarayana*[3]

[1]Research Center Imarat, Hyderabad
[2]Research Center Imarat, Hyderabad
[3]INSA Senior Scientist, IIIT-Hyderabad

BHVSNM@rcilab.in, satyanarayana.jv@rcilab.in, yegna@iiit.ac.in

Abstract

The objective of the study is to show that a speaker's location in a practical environment can be obtained from the time delays of the speech received at spatially distributed microphones. The time delay at a pair of microphones is estimated reliably using a recently proposed single frequency filtering (SFF) analysis of speech even when the speech collected in a live room is degraded due to echoes, reverberation and audio signals from other sources. The reliability is due to evidence of time delay from multiple frequency components obtained in the SFF analysis. The effectiveness of the proposed method for determining the speaker location can be demonstrated using a pair of microphones for picking up the speech signals, and then processing the signals using SFF analysis.

Index Terms: Time delay estimation, time lag, cross-correlation, single frequency filtering, speech signals

1. Introduction

Identification of location of each speaker amongst many sitting at fixed distances can aid special attention mechanisms that provide better audio-visual experience to the audience. The time lag between the received speech signals at two spatially distributed microphones is specific to each speaker, assuming that the speakers are nearly static at a given location. The lag estimation (LE) between signals arriving at two sensors is usually done using some form of cross-correlation function [1]. Several methods [2, 3] have been proposed for pre-processing the signals before computing the cross-correlation function to mitigate the effects due to multi-path, reverberation and noise. Weighting the crossspectrum has been studied extensively to overcome the effects of degradations due to noise and reverberation [4, 5, 6]. Several reviews [2, 7, 8] of lag estimations are available.

In this paper, a lag estimation method based on the recently proposed single frequency filtering (SFF) analysis of speech signals [9] is presented. Through this method, evidence for integer lag (in number of samples for a given sampling frequency) can be obtained, not only from successive segments of the signal, but also from the components of the signal at several frequencies [10]. In addition, the amplitudes of the correlation sequence can be used to obtain the true lag through fractional lag estimation (FLE)[10].

Section 2 gives a brief review of the SFF analysis for decomposition of a signal into number of components, each corresponding to one frequency. Lag estimation from SFF components is explained in section 3. Section 4 gives the steps in FLE and section 5 explains how FLE can be improved by including a neural network model in the design. Section 6 describes the experimental setup for demonstration of the proposed method of locating a speaker in a live environment.

2. Single Frequency Filtering

SFF facilitates decomposition of the signal into components at individual frequencies. The envelopes and the corresponding phases are obtained at any desired frequency by passing the frequency-shifted signal through a near ideal resonator located at $f_s/2$, where f_s is the sampling frequency. The steps involved in computing the SFF output at a given frequency f_k are as follows [9]:

1. The speech signal $s[n]$ is differenced to reduce any low frequency trend in the recorded signal.

$$x[n] = s[n] - s[n-1]. \tag{1}$$

2. The frequency-shifted signal is given by

$$x_k[n] = x[n]e^{j\bar{\omega}_k n} \tag{2}$$

where $\bar{\omega}_k = \pi - \omega_k$.

3. The signal $x_k[n]$ is passed through a single pole filter,

$$H(z) = \frac{1}{1 + rz^{-1}}. \tag{3}$$

To ensure stability of the filter, the value of r is chosen close to, but less than 1. In this work, $r = 0.99$ is used. The filtered output is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \tag{4}$$

4. The magnitude or envelope $v_k[n]$ and the phase $\theta_k[n]$ of the signal $y_k[n]$ are given by

$$v_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]} \tag{5}$$

and

$$\theta_k[n] = tan^{-1}(\frac{y_{ki}[n]}{y_{kr}[n]}) \tag{6}$$

respectively, where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary parts of $y_k[n]$.

Since speech samples are correlated and noise samples lack correlation, $v_k[n]$ has some high SNR regions, which can be made use of for lag estimation. Also, evidence for the lag is available at several frequencies in the SFF analysis. Effects due to waveform distortion are reduced in the SFF outputs, as we consider the envelope of the signal at each frequency separately.

3. Integer lag estimation using SFF (ILESFF)

The steps for computing ILESFF are as follows [10]:

1. For the time-delayed signals $x[n]$ and $y[n]$, the SFF envelopes at each of the K frequencies (separated by 10 Hz) are computed. For f_s=8000Hz, K=400.
2. For each utterance, there are N segments of the envelopes, each of size 50 msec with a shift of 5 msec. Thus there are $K \times N$ cross-correlation sequences, $R_{xy}[l]$ between corresponding segments of the SFF envelopes at each frequency.
3. For each segment, the average of the cross-correlation sequences across the K frequencies is obtained.
4. The lag corresponding to the peak of the average correlation sequence is found for each of the N segments. The estimated lag is the one which is associated with the correlation peak in maximum number of segments.

4. Fractional lag estimation using SFF (FLESFF)

Let $L \in I$ be the integer lag associated with the peak of the cross-correlation sequence $R_{xy}(l)$ for a given segment at a given frequency, i.e., $R_{xy}(L) > R_{xy}(l)$ for all l. Further, let $l_t \in \mathbb{R}$ be the true lag with a fractional part. It can be shown that [10]:

- Case (i): If $R_{xy}(L-1) - R_{xy}(L+1) > \theta$, then $L-1 < l_t < L$
- Case (ii): If $-\theta \leq R_{xy}(L-1) - R_{xy}(L+1) \leq \theta$, then $L \approx l_t$
- Case (iii): $R_{xy}(L-1) - R_{xy}(L+1) < -\theta$, then $L < l_t < L+1$

where θ is a small threshold, chosen as 0.03 in this study.

Following are the steps in the FLESFF algorithm [10]:

1. Initialize three counters: c_L, c_{L-1} and c_{L+1} to zero.
2. The cross-correlation sequences between the corresponding segments in the SFF envelopes for each frequency are computed as in ILESFF.
3. Identify the lag corresponding to the peak in each of the $K \times N$ correlation sequences.
4. In all those correlation sequences, where L is the lag of the peak, determine the relative position of the true lag with respect to L. Increment the counters, c_{L-1}, c_L and c_{L+1}, respectively, according to Case (i), Case (ii) or Case (iii).
5. Obtain an estimate of the fractional lag[1] as weighted average, i.e.,

$$\tilde{l}_t = \frac{(L-1)c_{L-1} + Lc_L + (L+1)c_{L+1}}{c_{L-1} + c_L + c_{L+1}} \quad (7)$$

5. Fractional lag estimation using artificial neural network (FLESFFNN)

Replacing the last step above with a feedforward neural network model for estimation of the true lag gives substantial improvement in accuracy [10].

The training data for the neural network has 10000 input-output pairs, each pair having

- Input vector of lag counts: $[\hat{c}_{L-1}, \hat{c}_L, \hat{c}_{L+1}]$ in which each element is normalized by division with $(c_{L-1} + c_L + c_{L+1})$.
- Output: $(l_t - L)$, the difference between ground truth and the integer lag

For lag computation, (7) in Section 4 is replaced by

$$\tilde{l}_t = f(W^*, \hat{c}_{L-1}, \hat{c}_L, \hat{c}_{L+1}), \quad (8)$$

where f represents the neural network regression function, and W^* is the weight vector of the trained network.

The lag calculated using FLESFFNN in a simulated setup was observed to be within an error of 0.2 from the true ground truth lag in 97% of the cases. Robust identification of speaker locations can be obtained using FLESFFNN.

6. Demonstration

Our experimental setup comprises a known number of speakers, seated at fixed locations near two microphones in a live room. The lag estimate, within a defined numerical range, at any instant of time uniquely identifies the speaker location and is highlighted on a display driven by a camera. The system is designed to ignore a lag caused by a new source, which may enter the system at locations other than the fixed locations.

This paper is meant to demonstrate the application of the lag estimation using SFF [10, 11], in a live room environment.

7. References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay", IEEE Trans. on Acoustics, Speech, and Signal Processing, 24(4), 320–327, Aug 1976

[2] J. Chen, J. Benesty and Y. Huang, "Time delay estimation in room acoustic environments: An overview", EURASIP J. Appl. Signal Process., 119, Jan. 2006

[3] J. Benesty, Y. Huang, "Time-delay estimation via linear interpolation and cross correlation", IEEE Trans. Speech and Audio Processing 12(5), 509-519, 2004

[4] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments", International Conference on Acoustics, Speech, and Signal Processing, 3021-3024, May, 2001

[5] S. Bulek and N.Erdol, "Effects of cross-spectrum estimation in convolutive blind source separation: A comparative study", Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE),122-127, Jan 2011

[6] B. Champagne, S. Bedard and A. Stephenne, "Performance of time delay estimation in the presence of room reverberation", IEEE Trans. on Speech and Audio Processing, 4(2), 148-152, Mar 1996

[7] A. Brutti, M. Omologo and P. Svaizer, "Comparison between dierent sound source localization techniques based on a real data collection", Hands-Free Speech Communication and Microphone Arrays (HSCMA), 69-72, May 2008.

[8] F. Wen and Q. Wan, "Robust time delay estimation for speech signals using information theory: A comparison study", EURASIP Journal on Audio, Speech, and Music Processing, 1-10, 2011

[9] G. Aneeja,and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and non-speech", IEEE/ACM Trans. on Audio, Speech, and Lang. Process., 23(4), 705-717, Apr. 2015

[10] B.H.V.S. Narayanamurthy, J.V. Satyanarayana and B. Yegnanarayana, "Fractional Time Delay Estimation from Broadband Signals like Speech", submitted to Interspeech 2018 as a regular paper

[11] B. Yegnanarayana, B.H.V.S. Narayanamurthy and Sudarsana Reddy Kadiri,"Single Frequency Filtering for Time Delay Estimation from Multispeaker Data",submitted to Interspeech 2018 as a regular paper

[1] Fractional lag, refers to a real number which may be greater than 1