INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Instantaneous Time Delay Estimation of Broadband Signals

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

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

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

Abstract

This paper presents a method of obtaining the instantaneous time delay of broadband signals collected at two spatially separated microphones in a live room. The method is based on using the complex signals at the output of single frequency filtering (SFF) of the microphone signals. We show that the complex SFF spectrum at each instant can be used to obtain the instantaneous time delay (TD). By using only the phase of the SFF spectrum, it is possible to get a better estimate of the TD, as in the case of the standard GCC-PHAT method. We show the effectiveness of the proposed method for real microphone signals collected in a live room. Robustness of the method is tested for additive babble noise at 0 dB for the live microphone data. Since we get the TD at every sampling instant, it may be possible to exploit this feature for two-channel multi-speaker separation and for tracking a moving speaker.

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

1. Introduction

The primary interest in analysing broadband signals like speech is in the estimation of the characteristics of the sources, i.e., the number of individual sources, source separation, source localization and tracking. Time delay estimation (TDE) is usually the first step in these studies [1]. The TDE involves estimating the time delay between the signals received at two or more microphones from a broadband source. The related problems associated with the TDE are:

- Direction of Arrival (DoA) estimation
- Tracking the locations of sources
- Estimation of the number of sources
- Separation of the individual sources

The time delay between two source signals $x[n]$ and $y[n]$ arriving at two different microphones is estimated using the cross-correlation function of the two signals, given by

$$R_{xy}(l) = \sum_{n=0}^{N-1} \{x[n]y[n+l]\}, \quad -(N-1), \leq l \leq (N-1), \tag{1}$$

where N is the number of samples. It is assumed that $x[n]$ and $y[n]$ are normalized signals, with root mean square (RMS) value of the signal being the normalizing factor.

The cross-correlation function in (1) is computed using the complex Fourier Transform (FT) of the signals, especially by weighting the product of the FTs with the inverse of the product of the magnitudeS of the FTs of the signals. This technique is known as the GCC-PHAT method [2]. A generalization of this approach [3, 4, 5, 6, 7] is to use a suitable weighting function in the spectral domain before computing the inverse FT. For broadband signals, the TD is estimated for each frequency band, and the results are combined. The TDE-based on higher order statistics was reported in [8, 9]. A survey of major contributions to TDE spanning over three decades is given in [1].

Sophisticated approaches for TDE are being explored by deriving a time-frequency (T-F) mask using deep neural networks (DNN), and then obtaining the cross-correlation function for the TDE. The authors in [10] investigate deep learning based time-frequency (T-F) masking for robust time difference of arrival (TDOA) estimation in noisy and reverberant environments. The key idea is to leverage the power of the DNN to determine the T-F units that are relatively clean for the TDOA estimation. In [11], the authors propose a neural network model for simultaneous detection and localization of multiple sound sources. They employ a likelihood-based encoding of the network output, which naturally allows the detection of an arbitrary number of sources. In addition, they investigate the use of sub-band cross-correlation information as features for better localization from sound mixtures, as well as three different network architectures based on different motivations. Good performance for TDE and acoustic source localization was reported using supervised training of neural networks in [12].

Recently, a new method based on single frequency filtering (SFF) of microphone signals is proposed for TDE [13]. The SFF method gives instantaneous complex SFF spectra. Using only the SFF magnitude signals, it was shown that the estimated TD gives performance better than the standard GCC-PHAT method for determining the number of speakers from multi-speaker mixed signals at two spatially separated microphones in a live room [13].

In this paper, we propose the use of the complex SFF spectrum to obtain the TD at each sampling instant. The instantaneous TD can be obtained from speech data collected at two spatially separated microphones in a live room. The paper is organized as follows. Section 2 gives the steps involved in the SFF decomposition of signals. Section 3 describes the proposed method for estimating the instantaneous TD, using the SFF spectra. Section 4 gives the results of the studies with two microphone signals. The key contributions of the paper are summarized in Section 5.

2. Single Frequency Filtering

Single Frequency Filtering facilitates decomposition of a signal into components at individual frequencies. The magnitude and the corresponding phase as a function of time are obtained at any desired frequency, by passing the frequency-shifted signal through a near ideal resonator, whose pole is located at half the

http://dx.doi.org/10.21437/Interspeech.2020-2462

sampling frequency $f_s/2$. The steps involved in computing the SFF output at a given frequency f_k are as follows [14]:

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

$$x[n] = s[n] - s[n-1]. \quad (2)$$

2. The frequency-shifted signal

$$x_k[n] = x[n]e^{j\bar{\omega}_k n} \quad (3)$$

is obtained by multiplying the differenced signal $x[n]$ with $e^{j\bar{\omega}_k n}$, where $\bar{\omega}_k = \pi - \omega_k = \pi - 2\pi f_k/f_s$.

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

$$H(z) = \frac{1}{1 + rz^{-1}}. \quad (4)$$

where the pole is located on the negative real axis close to the unit circle in the $z-$ plane. The value of $r = 0.995$ is used in this work, although the value is not critical. The filtered output $y_k[n]$ is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \quad (5)$$

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

$$e_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]} \quad (6)$$

and

$$\theta_k[n] = tan^{-1}(\frac{y_{ki}[n]}{y_{kr}[n]}), \quad (7)$$

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

Since speech samples are correlated and noise samples are less correlated, $e_k[n]$ has some high SNR regions [14], which can help in better estimation of the time delay. Also, the evidence for the time delay is available at several frequencies due to the SFF analysis. Effects due to waveform distortion can be reduced in the SFF outputs, if the signal component at each frequency is considered separately. In this study, the SFF outputs are extracted for f_k of every 10 Hz.

3. Instantaneous TDE from complex SFF spectra

The SFF magnitude and phase for all frequencies at any instant constitute the instantaneous complex SFF spectrum. Let $y_{1k}[n]$ and $y_{2k}[n]$ be the k^{th} component of the complex SFF spectra from the two microphone signals at the time instant n. Then

$$y_{1k}[n] = e_{1k}[n]e^{j\theta_{1k}[n]} \quad (8)$$

and

$$y_{2k}[n] = e_{2k}[n]e^{j\theta_{2k}[n]} \quad (9)$$

where suffixes 1 and 2 refer to microphone 1 and 2, respectively. The two SFF spectra $y_{1k}[n]$ and $y_{2k}[n]$, $k = 0, 1, ..., K-1$, are like the Fourier transforms (FT) of the segments of the signals around n. The inverse Fourier transform of the product $y_{1k}[n]y_{2k}^*[n]$, $k = 0, 1, ..., K-1$, gives a time sequence which is like cross-correlation sequence at that instant. The location

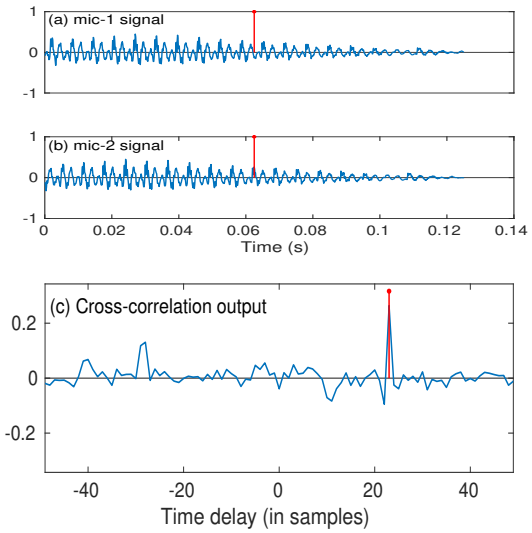

Figure 1: *(a) Waveform of microphone-1 signal. (b) Waveform of microphone-2 signal. (c) Cross-correlation output at a given instant (shown with 'red' stem in (a) and (b)). The 'red' stem in (c) indicates the location of maximum in the cross-correlation function.*

of the peak in the cross-correlation sequence corresponds to the instantaneous time delay between the signals received at the two microphones. The time delay values are limited to the integer number of time intervals obtained while computing the inverse Fourier transform of the product of the SFF spectra. The time resolution can be improved by increasing the number of appended zeros before computing the inverse Fourier transform.

Dividing the product $y_{1k}[n]y_{2k}^*[n]$ by the magnitude of the product gives the phase component of the product for all frequencies. That is, we get

$$\frac{y_{1k}[n]y_{2k}^*[n]}{|y_{1k}[n]y_{2k}^*[n]|} = e^{j(\theta_{1k}[n] - \theta_{2k}[n])}, k = 0, 1, ..., K-1 \quad (10)$$

This is like GCC-PHAT on signals [2]. The inverse Fourier transform of this sequence gives an output like a cross-correlation sequence, with the location of the peak corresponding to the time delay.

4. Results of TDE studies with two microphone signals

In this section, we discuss the results of studies made with two microphone signals. In the first study, we have estimated the time delay between two synthetically delayed signals. The test signal of a single speaker was taken from the TIMIT corpus [15]. A delayed version of the signal is created with a known integer delay, which serves as the ground truth for the experiment. In the second study, utterances from two speakers from the TIMIT corpus are added with different delays. In the third study, speech utterances from two speakers are recorded using two spatially separated microphones in a live room. Note that in this case, the ground truth of the time delay is not known. In the fourth study, speech uttered by a moving speaker is recorded by two spatially separated microphones in a live room.

Fig. 1(a) shows an utterance of a single speaker taken from TIMIT corpus which is considered as mic-1 signal. Fig. 1(b)

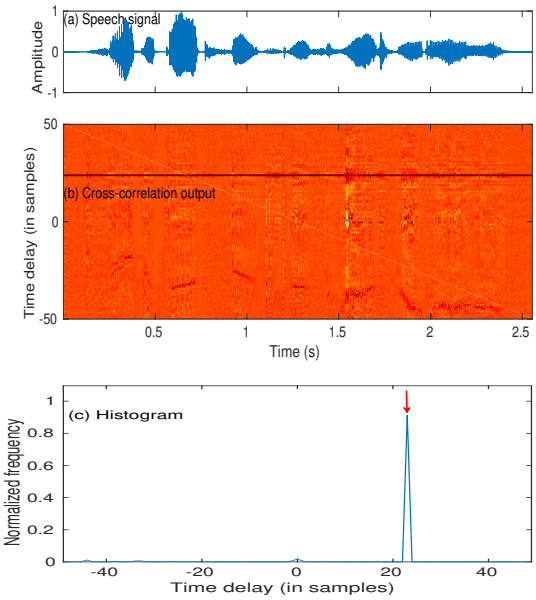

Figure 2: *(a) Waveform of microphone-1 signal for a single speaker, (b) Normalized cross-correlogram (cross-correlation function computed at each instant) is plotted using color reflecting the amplitude, and (c) histogram of the time delays obtained from the cross-correlation function given in (b).*

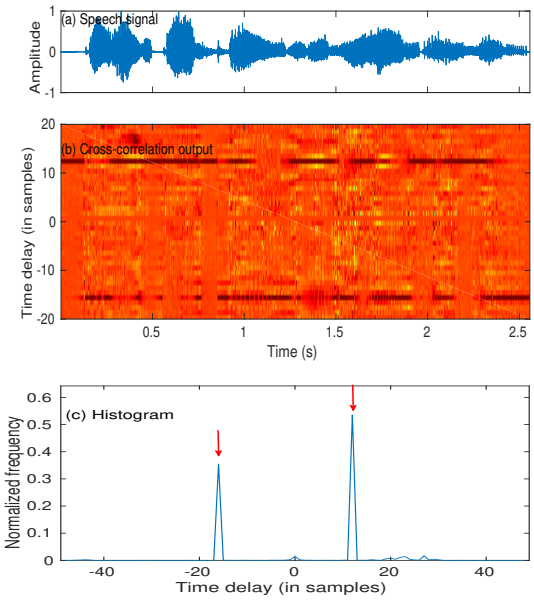

Figure 3: *(a) Waveform of microphone-1 signal, (b) Normalized cross-correlogram (cross-correlation computed at each instant) for two speaker case synthetically generated using TIMIT sentences, and (c) histogram of the time delays obtained from the cross-correlation function shown in (b).*

shows the same utterance in Fig. 1(a) delayed by 23 samples, and is considered as mic-2 signal. Fig. 1(c) shows the cross-correlation function computed at the given instant (shown by the red lines in Fig. 1(a) and Fig. 1(b)) for mic-1 and mic-2 signals, respectively. In Fig. 1(c), the location of the peak in the cross-correlation function is shown with a red line, which correspond to the time delay at that instant.

The time-delay is estimated at each instant using the cross-correlation function obtained from the complex SFF outputs as shown in in Eq (9). Fig. 2(a) shows the waveform of mic-1 signal and Fig. 2(b) shows the normalized cross-correlation function at each instant called normalized cross-correlogram, plotted using color reflecting the amplitude values. The cross-correlation function at each instant is normalized with the peak value before plotting as a cross-correlogram in Fig. 2(b). We can observe a constant line at the 23^{rd} sample, corresponding to the actual time-delay. A histogram of the estimated time delays is shown in Fig. 2(c). The location of the peak of the histogram is at the 23^{rd} sample, corresponding to the time-delay. The peak value of the histogram indicates that at more than 80% of the instants in the speech signal, the time-delay is estimated correctly, while at the other 20% of the instants the estimated time-delay is wrong due to low signal-to-noise ratio (SNR) or silence regions in the speech signal.

The time-delay estimation using the proposed method is also evaluated on synthetically mixed two speaker data. Two utterances from the TIMIT corpus are randomly selected, and are mixed to get mic-1 signal. In mic-2 signal, one of the utterance is delayed by -16 samples and the other utterance is delayed by 12 samples, and the delayed signals are mixed. Fig. 3(a) shows the mic-1 signal for the two speaker case, and 3(b) shows the normalized cross-correlation function at each instant (i.e., normalized cross-correlogram) plotted using color reflecting the amplitude values. Fig. 3(c) shows the histogram of the values in Fig. 3(b). The two strong peaks in the histogram indicate the presence of two speakers in the signal, and the location of the peaks at 12 and -16 indicate the delays between the microphones for each of the two speakers.

Fig. 4(a) shows the mic-1 signal for the two speaker case recorded in a live room, and Fig. 4(b) shows the normalized cross-correlation function at each instant (i.e., normalized cross-correlogram). Fig. 4(c) shows the histogram of the time delay values in Fig. 4(b). The two peaks in the histogram (pointed by red arrow) indicate the delays between the microphones for each of the two speakers.

Fig. 5(a) shows the mic-1 signal for a moving speaker case, recorded in a live room. Fig. 5(b) shows the normalized cross-correlation function at each instant (i.e., normalized cross-correlogram). The curved dark red line (starting around 0 samples and proceeding to negative delay) shows the trajectory of the time-delay of the moving speaker.

The robustness of the proposed instantaneous TDE is examined for the two speakers live data given in Fig. 4. Babble noise at 0 dB SNR (signal-to-noise ratio) is added to each microphone signal, and the resulting TDE plots are shown in Fig. 6. Fig. 6(b) shows the time delay indicated by dark red lines for the two speakers in the cross-correlogram plot. Fig. 6(c) shows the histogram plot indicating the presence of two speakers even at 0 dB SNR.

Fig.7 shows the cross-correlation values as a function of time at delays $T_1 = -7$ samples and $T_2 = 6$ samples for the two speakers live data given in Fig. 4. These values indicate the time intervals where each of the speakers are present. Note that there are some overlapping regions, indicating the overlapping speech of two speakers. This feature of instantaneous time delay may help in separating the speech of individual speakers

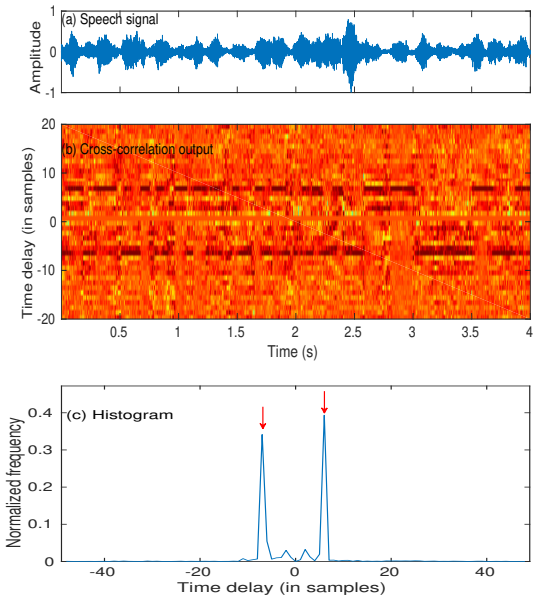

Figure 4: *(a) Waveform of microphone-1 signal, (b) Normalized cross-correlogram (cross-correlation computed at each instant) for two speaker case recorded in live room, and (c) histogram of time delays obtained from the cross-correlogram shown in (b).*

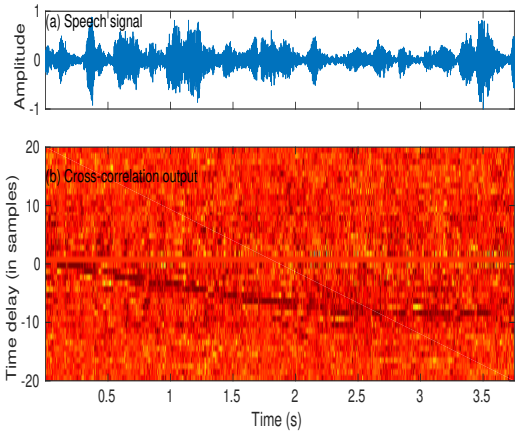

Figure 5: *(a) Waveform of microphone-1 signal, (b) Normalized cross-correlogram (cross-correlation computed at each instant) for a moving speaker recorded in a live room.*

from the two microphone mixed signals.

5. Conclusion

In this paper, a method for computing the instantaneous TD between speech signals, received at two spatially separated microphones, was proposed. This was possible because of the SFF decomposition of the signals, which makes the signal component available at any desired frequency. While the studies reported in [13] make use of only the magnitude of the SFF spectrum, the method proposed in this paper extends the idea to include the SFF phase. This allows for the computation of the instantaneous time delay. To the author's knowledge, this is the

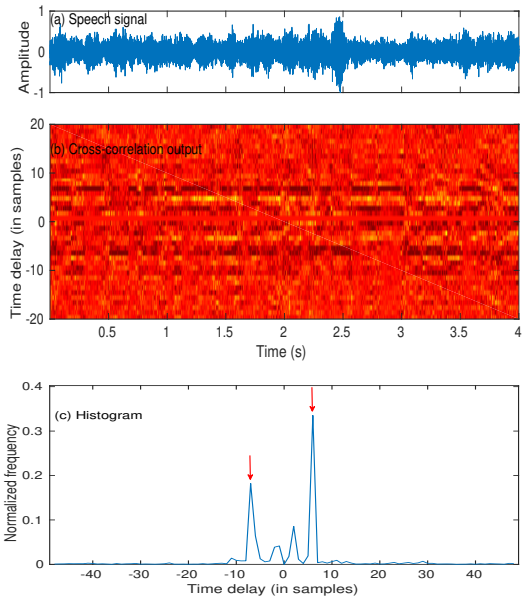

Figure 6: *(a) Waveform of microphone-1 signal, (b) Normalized cross-correlogram (cross-correlation computed at each instant) for two speaker case recorded in live room degraded by babble noise at 0 dB SNR, and (c) histogram of time delays obtained from the cross-correlogram shown in (b).*

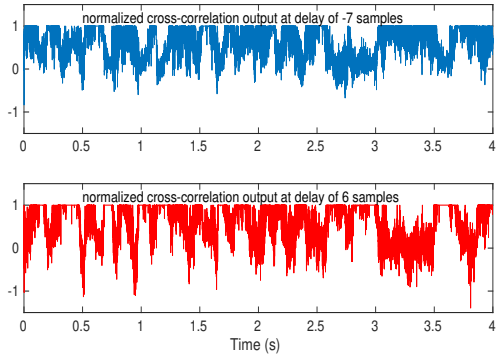

Figure 7: *Normalized cross-correlation output at delay of -7 samples and 6 samples corresponding to two peaks (two speakers) in histogram shown in Fig. 4(c).*

first time a method for obtaining the time delay at every sampling instant is proposed for data collected in a live room, while the speakers are speaking simultaneously. With the availability of the instantaneous time delay, the studies can be extended for multi-speaker separation and also for tracking the movement of a speaker in a live room. It is also possible to track multiple speakers by tracking multiple peaks, as the time delay due to each speaker will be distinct. But it is applicable under the constraint that the speakers and the microphones are approximately in the same plane, as we use only two microphones to collect the data. Since the time delay is directly controlled by the phase, this study may be extended for TDE by making use of only the phase information in the SFF spectra of the two microphone signals.

6. References

[1] J.Chen, J.Benesty, and Y.Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–19, 2006.

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[3] J. C. Hassab and R. E. Boucher, "Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 549–555, 1981.

[4] L. E. Miller and J. S. Lee, "Error analysis of time delay estimation using a finite integration time correlator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 490–496, 1981.

[5] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 998–1003, 1982.

[6] M. Azaria and D. Hertz, "Time delay estimation by generalized cross- correlation methods," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 280–285, 1984.

[7] Y. Bar-Shalom, F. Palmieri, A. Kumar, and H. M. Shertukde, "Analysis of wide-band cross correlation for time-delay estimation," *IEEE Transactions on Signal Processing*, vol. 41, no. 1, pp. 385–387, 1993.

[8] J. K. Tugnait, "Time delay estimation with unknown spatially correlated gaussian noise," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 549–558, 1993.

[9] Y. Wu, "Time delay estimation of non-Gaussian signal in unknown gaussian noises using third-order cumulants," *Electronics Letters*, vol. 38, no. 16, pp. 930–931, 2002.

[10] Z. Wang, X. Zhang, and D. Wang, "Robust tdoa estimation based on time-frequency masking and deep neural networks," *Proceedings of Interspeech 2018, Hyderabad, India*, pp. 322–326, Sep. 2018.

[11] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," *Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia*, pp. 74–79, 2018.

[12] L. Houegnigan, P. Safari, C. Nadeu, M. van der Schhaar, M. Sole, and M. Andre, "Neural networks for high performance time delay estimation and acoustic source localization," pp. 137–146, 2017.

[13] B. N. Murthy, B. Yegnanarayana, and S. Kadiri, "Time delay estimation from mixed multispeaker speech signals using single frequency filtering," *Circuits Syst. Signal Process (2019)*, Aug 2019.

[14] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discrimnating speech and non–speech," *IEEE/ACM Transaction on Audio, Speech and Language Processing*, no. 4, pp. 705–717, 2005.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.