INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Locating Burst Onsets using SFF Envelope and Phase Information

Bhanu Teja Nellore, RaviShankar Prasad, Sudarsana Reddy Kadiri, Suryakanth V. Gangashetty and B. Yegnanarayana

Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India
{bhanu.nellore, ravishankar.prasad, sudarsanareddy.kadiri}@research.iiit.ac.in,
{svg, yegna}@iiit.ac.in

Abstract

Bursts are produced by closing the oral tract at a place of articulation and suddenly releasing the acoustic energy built-up behind the closure in the tract. The release of energy is an impulse-like behavior, and it is followed by a short duration of frication. The burst release is short and mostly weak in nature (compared to sonorant sounds), thus making it difficult to detect its presence in continuous speech. This paper attempts to identify burst onsets based on parameters derived from single frequency filtering (SFF) analysis of speech signals. The SFF envelope and phase information give good spectral and temporal resolutions of certain features of the signal. Signal reconstructed from the SFF phase information is shown to be useful in locating burst onsets. Entropy and spectral distance parameters from the SFF spectral envelopes are used to refine the burst onset candidate set. The identified burst onset locations are compared with manual annotations in the TIMIT database.

Index Terms: stop sound, burst onset, single frequency filtering, phase reconstruction

1. Introduction

Detection of burst onsets is an important problem in speech processing. Burst onset detection improves performance of automatic speech recognition (ASR) systems [1–5], and aids in improving speech intelligibility [6]. Identification of burst onsets can also be used to refine the manner hypotheses of phone recognizer [7]. Furthermore, information on burst onsets aids in the determination of place of articulation of stop sounds [8] and in the computation of the voice onset time (VOT) [9].

Burst onset is a short duration phenomenon occurring in speech signals. A burst onset landmark is characterized by a closure inside the vocal tract, followed by a sudden release of the acoustic pressure built up during the closure duration [10]. The abrupt release of the acoustic pressure at the burst onset gives rise to a stream of broadband energy in the speech signal. A burst onset is therefore always followed by a small duration of frication and/or aspiration.

The characteristics of burst onsets vary depending on several factors such as phonation state, participating articulators, and place of articulation. Depending on these factors, a burst onset may be strong or weak. For example, in the case of voiced burst release, the pressure build up during the closure duration is low due to the presence of glottal activity. Consequently, the burst onset following may not be prominent. In either case, burst onset has significantly lower signal energy in a speech utterance compared to sonorant sounds. Burst onsets are also influenced by contexts. Burst onsets occuring at syllable endings may not have prominent energies [11]. The energy distribution across the burst regions exhibits behavior similar to noise, and hence detecting burst onsets in low signal to noise ratio (SNR) regions is a challenging problem.

Several methods exploit spectral and temporal characteristics for burst onset detection and burst characterization in terms of place of articulation. A temporal measure called Rate of Rise (RoR) extracted across different frequency bands along with a rule-based algorithm to identify burst onsets was proposed in [2]. Degree of abruptness, i.e., energy difference between two appropriately located frames is used as an acoustic measure in [3]. Support Vector Machine (SVM) with a three dimensional vector input was shown to perform better in detecting stops than Hidden Markov Model (HMM) based systems [12]. The three dimensional vector consists of parameters such as total energy of the signal, band energy above 3 kHz, and a measure of spectral flatness based on Wiener entropy. A method using Recurrent Neural Networks (RNNs) to detect burst onsets with standard frame-based spectral features was proposed in [4]. A set of spectral and temporal features like energy ratios and zero crossings were proposed for burst onset detection in [5]. Another method utilizes parameters extracted from the log magnitude spectrum, voicing onset offset and spectral flatness information to detect stop landmarks [13]. Parameters based on the rate of change of spectral moments along with spectral energy parameters are used for improving the performance of burst onset detection in [14]. A recent method [15] introduced a temporal measure called plosion index for the identification of burst onset in continuous speech.

Most of the methods for burst detection use spectral characteristics in different frequency bands, such as the spectral energy, or features relating to spectral shape such as spectral flatness. Other methods exploit the temporal behavior of burst onsets such as the sudden rise of energy after the period of closure. All these methods use discrete Fourier transform (DFT) based spectral representation, which employ block processing to estimate the characteristics in frequency domain. The block processing method assumes stationarity of the signal over the duration of analysis. Such an assumption may mask the transient nature of the burst. Further, it may lead to deviation of detected burst onset locations from the actual ones in the speech signal. This study focusses on locating burst onsets in continuous speech.

This study identifies burst onset locations, based on a recently proposed speech signal analysis method, called single frequency filtering (SFF) [16]. The SFF method provides temporal envelope of the speech signal at any desired frequency. The SFF method also gives instantaneous spectra. The present study uses the envelope and phase information of the SFF to determine the location of burst onsets. The signal reconstructed from phase information helps in emphasizing the burst onsets.

The paper is organised as follows. Section 2 discusses the

http://dx.doi.org/10.21437/Interspeech.2017-1027

significance of the features extracted using SFF method. Section 3 describes the experimental details and results. Section 4 gives a summary and conclusion of this study.

2. SFF parameters for the study of burst onsets

2.1. SFF envelope and phase reconstructed signal

The objective in SFF is to derive the amplitude envelope of the signal as a function of time at a desired frequency. The SFF analysis is performed using a resonator at $f_s/2$ (f_s = sampling frequency) for each frequency component, after frequency shifting the signal. This ensures that the filter characteristics remains same for every frequency component. Since the SFF is performed using a resonator at $f_s/2$, whose pole is located close to the unit circle, the effect of other frequency components are reduced significantly. Following are the steps involved in obtaining the amplitude envelope and phase information of the signal at the k^{th} desired frequency component (f_k Hz) [16]:

- The speech signal $s[n]$ is frequency shifted by $\bar{f}_k$, where $\bar{f}_k$ = ($f_s/2$)-f_k. The resulting frequency shifted version of the signal is given by

$$x_k[n] = s[n]e^{-jn\frac{2\pi\bar{f}_k}{f_s}}, \quad (1)$$

  for n = 1, 2, ..., N, and k = 1, 2, ..., K, where N is the total number of samples in the signal, and K is the total number of frequency components. If the filters are placed at every 1 Hz spacing, the value of K is $f_s/2$ in Hz.
- The signal $x_k[n]$ is passed through a single pole filter whose transfer function is given by $H(z) = \frac{1}{1+rz^{-1}}$. The pole of this filter is near to the unit circle ($r \approx 1$) on negative real axis in the z-plane.
- The output of the filter is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \quad (2)$$

  It is to be noted that $y_k[n]$ is a complex number with real part $y_{kr}[n]$ and imaginary part $y_{ki}[n]$.
- The envelope of the signal $y_k[n]$ is given by

$$e_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]} \quad (3)$$

  where $e_k[n]$ is the SFF envelope at the k^{th} frequency.
- The phase of the signal $y_k[n]$ is given by

$$\phi_k[n] = \tan^{-1}\left(\frac{y_{ki}[n]}{y_{kr}[n]}\right) \quad (4)$$

To ensure the stability of the filter, the value of r is made less than 1. The envelopes of the signal are obtained for different frequencies. For the purpose of analysis of burst onsets, we chose the frequencies at 10 Hz intervals, using r = 0.995.

In order to obtain phase reconstructed signal, the envelope values $e_k[n]$ are set to one and SFF phase is used to reconstruct the speech signal. SFF phase reconstructed signal $\tilde{x}[n]$ highlights burst characteristics. Fig. 1(a) shows a speech signal $s[n]$ and the signal reconstructed using the phase information $\tilde{x}[n]$ is shown in Fig. 1(b) respectively. The burst onsets occur at 250, 330 and 540 ms, in the figure, among which the segment occurring at 540 ms is a weak burst. The reconstructed signal $\tilde{x}[n]$

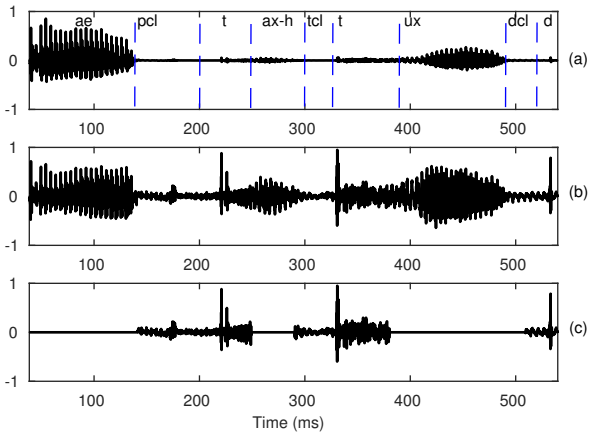

Figure 1: *(a) Speech signal for the utterance 'aptitude' (phone boundaries are marked using dashed line). (b) Signal reconstructed using phase only information. (c) Signal in (b) (only unvoiced regions are shown).*

highlights the burst behaviour of these regions as high amplitude transient behavior. Fig. 1(c) shows the reconstructed signal $\tilde{x}[n]$ in the unvoiced regions. An advantage of SFF phase is it's ability to reconstruct the bursts irrespective of their strength.

2.2. Spectral parameters from SFF envelope

The signal $\tilde{x}[n]$ can be used to highlight the discontinuities in speech. The discontinuities show up as high energy impulse-like behavior when reconstructed using the phase information alone. The spectral distance and entropy parameters derived from the envelopes $e_k[n]$ are further utilized to highlight the magnitude of spectral change around the regions of discontinuities in the signal. A burst onset location exhibits a significant increase in the spectral distance and also in spectral entropy.

Fig. 2 shows the speech signal corresponding to the utterance *'aptitude'* along with the temporal and spectral evidences obtained using the SFF method. Fig. 2(a) shows the speech signal along with the phone label in the utterance. The locations of the burst onsets corresponding to /t/, /b/ and /d/ can be demarcated in the signal at 250, 330 and 540 ms, respectively. Fig. 2(b) shows the signal reconstructed for the unvoiced region with a unit magnitude and the phase information $\phi_k[n]$. A significant impulse-like behavior around the burst onset can be noted in the signal. There are, however, other impulses in the reconstructed signal. Fig. 2(c) shows the gradient of squared Euclidean distance between successive SFF spectra. The squared Euclidean distance ($d_E[n]$) is given by

$$d_E[n] = \frac{1}{K}\sum_{k=1}^{K}(e_k[n] - e_k[n-1])^2, \quad (5)$$

where $e_k[n]$ and $e_k[n$-$1]$ are the SFF envelopes computed at the sampling instants n and n-1, respectively. A significant rise in the spectral distance can be seen in the figure around the burst onsets. Fig. 2(d) shows the differenced Wiener entropy ($e_W[n] - e_W[n-1]$) computed over $e_k[n]$ and $e_k[n-1]$. Wiener entropy $e_W[n]$ is given by,

$$e_W[n] = \frac{\frac{1}{K}\sum_{k=0}^{K-1}\ln e_k[n]}{\frac{1}{K}\sum_{k=0}^{K-1} e_k[n]}, \quad (6)$$

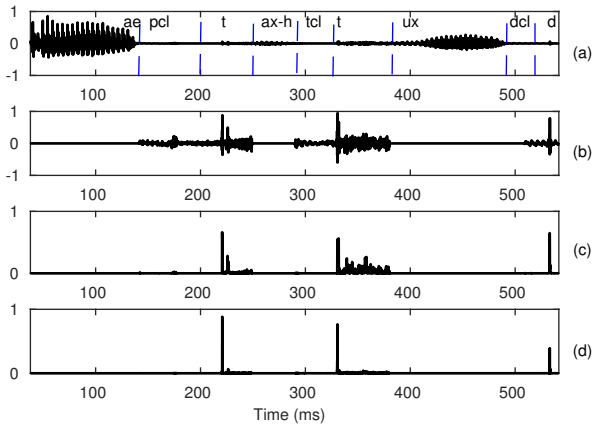

Figure 2: *(a) Speech signal. (b) Signal reconstructed using phase only information in unvoiced regions. (c) Difference of* $d_E[n]$*. (d) Difference of* $e_W[n]$*.*

where K are the number of frequency bins. The sudden rise in the $e_W[n]$ contour is captured in the differenced signal, for regions around the burst onsets. We use these envelope-based parameters as evidences to locate the burst onset with accuracy.

3. Study of burst onsets in TIMIT

3.1. Experimental details and proposed locations of burst onsets

The burst onset candidate set is identified around the instants of discontinuity in the signal $\tilde{x}[n]$. The spectral parameters obtained from SFF envelopes help in providing multiple evidences for refining the candidate set for the proposed burst onset locations. The behavior of these parameters are studied on clean speech as well as speech degraded with white noise at different SNRs. The proposed burst onsets are compared with the manually annotated burst onsets for the deviation between these locations, in TIMIT dataset [17]. A subset of TIMIT dataset consisting a total of 1000 utterances is considered in this study. The present study examines the burst sounds {/p/, /k/, /t/, /d/, /b/ and /g/}. The manual annotations demarcate the burst onset after the closure region. The present analysis uses the burst region along with closure region, denoted as R_b, to identify the burst onset location in the signal. A value of $r = 0.995$ is chosen for the SFF analysis, and the $\tilde{x}[n]$ signal is reconstructed from $\phi[n]$. The impulse–like discontinuities are identified in $\tilde{x}[n]$ within R_b corresponding to each annotated burst location, which constitute the candidate set. The release of acoustic energy may happen in more than one burst leading to a possibility of multiple candidates [15]. These bursts locations in $\tilde{x}[n]$, identified in the region R_b, constitute the candidate set $P_S = \{p1, p2, \ldots p_k\}$. To refine the burst onset candidate set, we further utilize the parameters $d_E[n]$ and $e_W[n]$ computed from the SFF envelopes. The behavior of these parameters are examined in a region of 1 ms around each candidate peak in P_S. The peak location in $\tilde{x}[n]$ with maximum cumulative change in the $d_E[n]$ and $e_W[n]$ parameters is chosen as the burst onset location.

3.2. Analysis on clean and noisy speech

The study uses the TIMIT annotations to demarcate the region R_b corresponding to each burst onset. In certain cases, the annotated burst onset location might not correspond to the location of discontinuity in the signal. We utilize the proposed parameters to highlight the location of significant discontinuity in the signal hypothesized as the correct burst onset locations, in the TIMIT dataset. Fig. 2 shows the behavior of the parameters $d_E[n]$ and $e_W[n]$ in the region R_b for different bursts occurring in the utterance 'aptitude'. We compare the annotated burst onset locations with the burst onset locations identified using proposed parameters. Fig. 3 shows histograms of the deviations obtained for different bursts. Figs. 3(a–g) present the population density vs. deviation (in ms) for the annotated burst onset locations vs. those obtained using proposed parameters. Fig. 3(a) shows population density for all the burst onsets. The histogram is skewed towards the right with a sharp peak in the 0–5 ms range, which covers almost 50% of the population. This suggests a small deviation in the manual annotation vs. discontinuities appearing in the signal. There is also a spread (almost 40% of the population) appearing uniformly across larger values of deviation. Figs. 3(b–g) present the histograms for unvoiced bursts /p/, /t/, /k/ and voiced bursts /b/, /d/, /g/ respectively. As seen in the figures, the voiced bursts exhibit more deviation than their unvoiced counterparts. This behavior can be attributed to the fact that voicing tends to a lower the amount of pressure build–up during the closure than their unvoiced counterparts. This results in relatively stronger burst release in voiceless cases which manifests as a stronger impulse in the signal [15]. Thus the voiceless bursts have lower percentage of deviation in the manual demarcations and the signal discontinuity (Figs. 3(b–d)), as these can easily be located. A larger population of voiced burst, on the other hand, exhibits a larger value deviation for a significant amount of bursts.

Table 1: *Percentage of burst onsets identified within a deviation of* 10 *ms at different SNRs.*

	All Bursts	/p/	/t/	/k/	/b/	/d/	/g/
Clean	79.2	86.1	75.9	84.3	69.9	79.6	74.1
40 dB	78.5	82.8	75.9	82.8	69.9	80.5	74.1
30 dB	74.6	73.0	73.2	79.3	67.7	79.6	69.0
20 dB	65.6	59.8	66.5	69.2	59.1	69.9	63.8

With respect to the place of articulation of bursts, velar stops /k/ and /g/ (Figs. 3(d) and 3(g)) show a higher population with less deviations for both unvoiced and voiced cases, respectively. This behavior can be attributed to a smaller area of contact at the place of articulation for velar stops resulting in a sharp release of the acoustic pressure [10, ch. 7]. Alveolar stops /t/ and /d/ (Figs. 3(c) and 3(f)) on the other hand have relatively more area of contact at the place of articulation [10, ch. 7]. This behavior results in a higher population of these bursts with more deviation. Among the velar bursts, the voiced burst /d/ exhibits larger deviation than voiceless counterpart /t/.

The bilabial bursts /p/ and /b/ (Figs. 3(b) and 3(e)) show distinct behavior with change in phonation state. The voiceless burst /p/ exhibits a larger population with lesser deviation. However, in the voiced case, due to the lesser pressure build up and a weak release, a larger population exhibits higher values of deviation [15]. Hence it becomes difficult to identify /b/ onset location.

The proposed method of locating the burst onsets is also evaluated for degraded speech signals. Fig. 4 shows the results for the population density vs. deviation for speech added

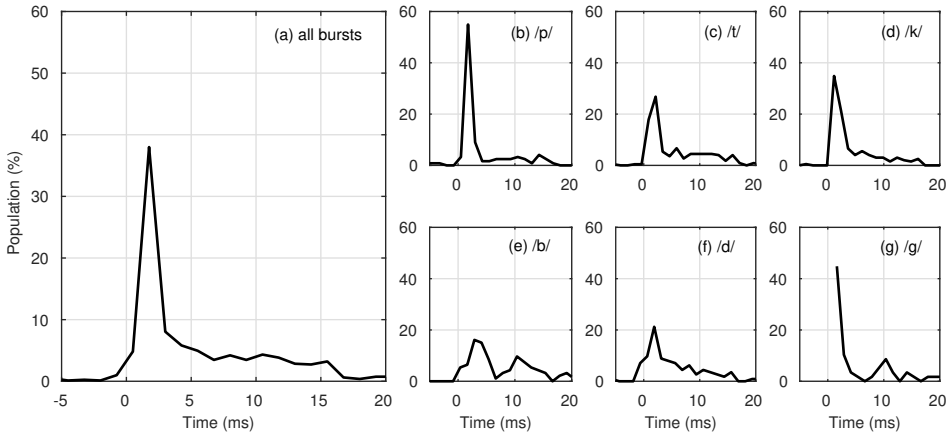

Figure 3: *Histogram distribution for deviation in proposed burst onsets vs. manual annotations (a) all bursts, (b)-(g) /p/, /t/, /k/, /b/, /d/ and /g/, respectively.*

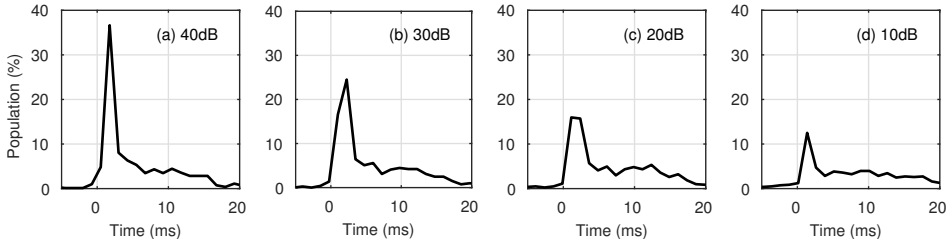

Figure 4: *Histogram distribution for deviation in manual annotations vs proposed burst onset, for speech added with white noise at (a)* 40*, (b)* 30*, (c)* 20 *and (d)* 10 *dB SNR, respectively.*

with white noise, at different SNRs. A comparison between results obtained for clean speech (Fig. 3(a)) and at SNR levels 40 dB and 30 dB (Figs. 4(a) and 4(b)) shows the performance of the method based on proposed parameters is robust to white noise up to 30 dB SNR. For SNRs lower than 20 dB, the performance drops significantly as the resolution of the impulses at the discontinuities in $\tilde{x}[n]$ degrades. This can also be observed from Table 1, which shows the burst onsets (in %) for different stop sounds, identified within a deviation of less than 10 ms, for cases of clean speech and speech corrupted with white noise across different SNRs. We can also observe that there is a large reduction in the population density of burst locations from 30 dB to 20 dB SNR.

4. Summary and conclusion

In this study, burst onsets corresponding to different stop consonants (/p/, /t/, /k/, /b/, /d/ and /g/) are analyzed. The SFF method is used to extract the features from the signal, which helps in highlighting the discontinuities in the signal. A set of burst candidates are extracted within a closed region, using phase information based reconstructed signal, and parameters based on gradients in Euclidean distance and Wiener entropy from the instantaneous spectra. Burst onset locations determined from the proposed parameters are compared with their manually annotated locations. Study highlights a close correspondence between the two factors for a significant number of examples from TIMIT dataset. The analysis highlights the importance of factors such as voicing and the area of contact influencing the demarcation of burst onset in speech. Among all the burst, the study suggests that the location of velar voiceless bursts are demarcated with the highest accuracy.

5. Acknowledgments

The third author would like to thank Tata Consultancy Services (TCS), India for supporting his PhD program.

6. References

[1] C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1253–1264, 2011.

[2] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

[3] N. N. Bitar, "Acoustic analysis and modeling of speech based on phonetic features," Ph.D. dissertation, Boston University, Massachusetts, 1998.

[4] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[5] J. Hou, L. Rabiner, and S. Dusan, "Automatic speech attribute transcription (ASAT)-the front end processor," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. I–333–I–336.

[6] A. R. Jayan and P. C. Pandey, "Automated modification of consonant-vowel ratio of stops for improving speech intelligibility," *International Journal of Speech Technology*, vol. 18, no. 1, pp. 113–130, 2015.

[7] N. Dhananjaya, B. Yegnanarayana, and V. Suryakanth, "Acoustic-phonetic information from excitation source for refining manner hypotheses of a phone recognizer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5252–5255.

[8] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358–1368, 1978.

[9] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3965–3979, 2012.

[10] K. N. Stevens, *Acoustic phonetics*. Massachusetts: MIT press, 2000.

[11] P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," *the Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1063–1076, 2002.

[12] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Arizona, U.S.A., Mar. 1999, pp. 425–428.

[13] A. R. Jayan and P. C. Pandey, "Detection of stop landmarks using gaussian mixture modeling of speech spectrum," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 4681–4684.

[14] A. R. Jayan, P. S. R. Bhat, and P. C. Pandey, "Detection of burst onset landmarks in speech using rate of change of spectral moments," in *Proce. National Conference on Communications (NCC)*, Bengaluru, India, Jan. 2011, pp. 1–5.

[15] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.

[16] 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, 2015.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.