



(RESEARCH ARTICLE)



## Analysis of vowel addition or deletion in Continuous Speech

Balakrishnan Sivakumar<sup>1</sup> and Praveen Kadakola Biligirirangaiah<sup>2,\*</sup>

<sup>1</sup> Department of ETE, Dr. Ambedkar Institute of Technology, Bangalore, India.

<sup>2</sup> Department of Electronics Engg, PRIST University, Thanjavur, India.

Global Journal of Engineering and Technology Advances, 2021, 07(03), 136–143

Publication history: Received on 08 May 2021; revised on 13 June 2021; accepted on 16 June 2021

Article DOI: <https://doi.org/10.30574/gjeta.2021.7.3.0084>

### Abstract

In order to improve the recognition performance, the articulation of the transcription is very important in the process of training. For continuous speech, the essential characteristics of various speakers are pronunciation variation, over focused or inadequately highlighted words can results the waveform misalignment in the sub word unit margin. Because of the deviation in the articulation leads into misalignment when this is compared with articulation dictionary. So the deletion or insertion of the sub word is necessary. This happens because for each expression, the transcription is not precise. This paper presents the corrections in the transcription at the sub word level utilizing sound prompts that are presented in the waveform. The transcription of a word is fixed Utilizing sentence-level transcriptions with reference to the phonemes that create the word. Specifically, it clarifies that vowels are either deleted or inserted. To help the proposed contention, errors in persistent discourse are validated utilizing machine learning and signal processing tools. A programmed information driven annotator abusing the inductions drawn from the examination is utilized to address transcription errors. The outcomes show that rectified pronunciations lead to higher probability for train expressions in the TIMIT corpus.

**Keywords:** data driven annotation; Speech transcription; Acoustic cues; Pronunciation variability

### 1. Introduction

The speech sounds are generally divided into two main types: vowels and consonants. Vowels are usually associated with high energy and strong time intervals. The comparative importance of vowels and consonants in conversational comprehension has been the subject of many studies. In studies that use colloquial sentences in the absence of background sound, vowels play an important role in recognizing words rather than consonants [1]–[3]. In the existence of noise, the vowels transmit on more speech information, because pattern notes may occur even louder in sound [4].

Most Indian languages have alphabet letters which resemble and teach users of related sounds. But in some languages, as in English, there is a lack of correspondence between the letters of the alphabet and the sound in which they stand. English writing styles never ensure proper pronunciation. This disadvantage of English produced the need of extra images that may speak to every particular English sound unmistakably. Accordingly, International Phonetic Association (IPA) gave the IPA Alphabet to the clients of English.

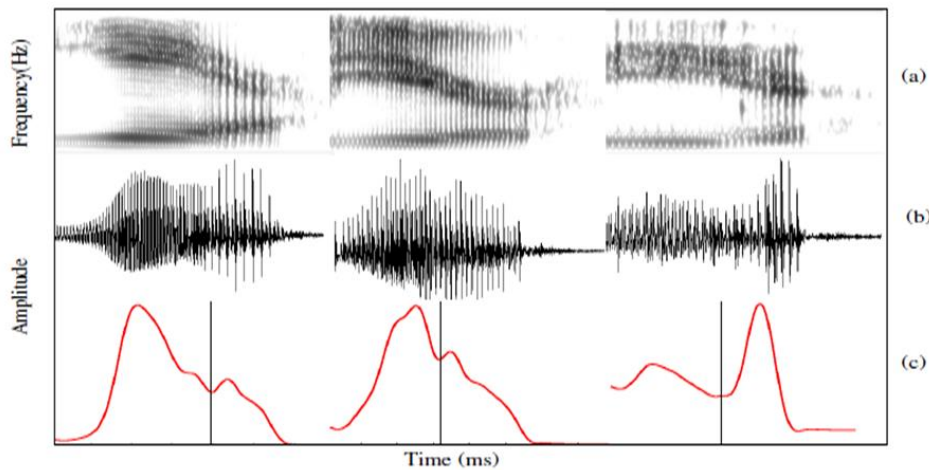
The most complex problem of awareness faced by non-native speakers is the placement of nuclear tones and the direction in which they should go. His speakers consciously use the correct awareness method; at the same time, non-native speakers need to study it based on certain rules, or in some cases most refer to the underlying form of their own mother tongue in English. Such implementation or variation in the use of the synchronization form gives the language a unique color.

\*Corresponding author: Praveen KB  
Department of Electronics Engg, PRIST University, Thanjavur, India.

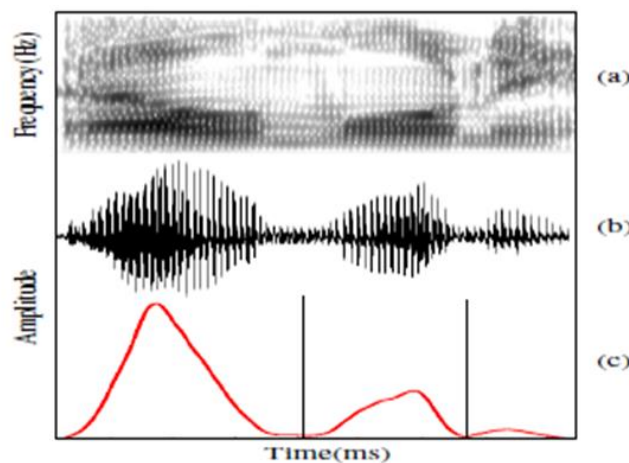
Developing a speech recognition system requires an understanding Conversational comment. First, it needs to be understood Speech units visible to the human ear. That's it literature has found a closeness to the characters link to humanoid language awareness and performance [5].The syllable is the smallest unit of speech production. Many psycholinguistic studies explored the role of syllable units in speech production and several offline studies conversational writing is action writing process [6]. Greenberg [7] opinions to that investigation as well the pronunciation variation at the syllable level is very organized.

Van Bale and others [8] have shown Automatic speech authentication system effectiveness with automatic transcriptions physically confirmed transcripts are comparable. Nevertheless, studies of articulation modeling [9, 10] demonstrate the same significance of orthographic record words in the environment of speech identification systems. Includes variation in pronunciation in speech identification systems by adding alternate accents dictionary of pronunciation. No attempt was made add these accents to sentence level orthography transcription of wave files. On this paper, considering the wave file and its transcription, we suggest a change Transcription using audio notes obtained using Signal processing equipment. Used to write unit research.

The language selected for the study is pure English Environment. English is a stress time language, period is one of the nearest pressure units. According to nearby distance compressed units can be either added or removed. In specific, deleted or added units are vowels.



**Figure 1** Extra syllable is resulted due to addition of new vowel (a) Spectrogram (b) Waveform (c) Smoothed energy



**Figure 2** Vowel deletion resulting in loose of a Syllable. (a) Spectrogram (b) Waveform (c) Smoothed energy

In this paper, we indicate deviations from the recommended form Continuous / Spontaneous speech through the CMU dictionary [11]. For example, pronunciation the term "year" recommended by the CMU vocabulary [12] is "y ih r". Intermittently, here is an additional pressure. It can happen adding a new vowel "er" to "y iher" monosyllabic is caused by the presence of the extra life of "er" The word CVC form becomes a bisyllabic term CVVC format is demonstrated in Figure 1. The character is classified as "a unit consisting of a rime and a coda, which reach the limit with the greatest vibration, while the vitality decreases to the beginning and the coda" [13]. Figure 1 shows an explanation of the three variants of the word year. Energy draws a boundary where energy decreases. It refers to the addition of vowels. As it can be shown in waves, spectrograms and energy plot. The peaks of the two main reserve energy plots indicate the existence of two characters. Pronunciation can sometimes lead to the removal of adverbs as well. For eg, Fig 2 indicates the cancellation of the vowel "ah" in "aelah cut niy".

The term comprises of four syllables, four protuberances are normal in the vitality shape. Three protuberances are demonstrated in figure 2. The elocution for this expression is "ael cut niy". A syllable fragment is in this manner erased. The perpendicular bars in all the Figures are conceivable syllable restrictions Ease of Use.

The proposed method tries to detect automatically variations in the pronunciation using transcript-level transcriptions. The mispronunciations location can be determined by using machine learning and signal processing tools.

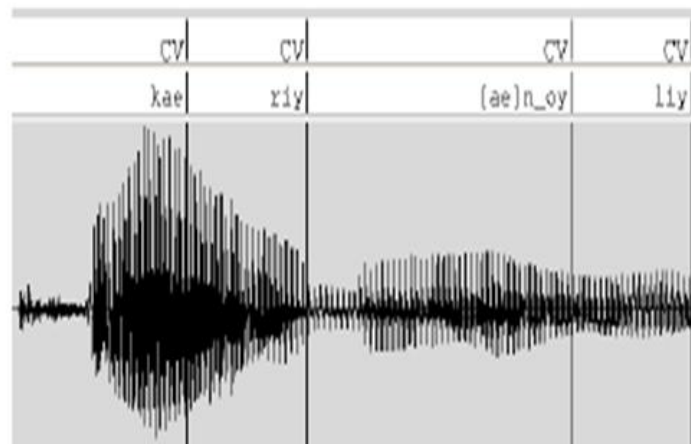
The variations in the Pronunciation [14] was examined in the context of the characters. The error statistics of telephone based contextual speech recognition is carried out using postmortem analysis In contrast to this study, an empirical study [15] showed the effect of pronunciation variation in speech recognition. Voluntary Accents used during recognition and improves performance in small vocabulary. [16] Like this sheet it also analyzes the pronunciation variation in the composition. In order to correct the transcription of the training data we are going to utilized a group delay methodology in consonance with Viterbi forced alignment and MLP based silence-vowel Consonant detection.

## 2. Overview of the tools used

There are three different tools have been utilized in order to automate the process of transcription misalignment identification in the waveform about the syllable.

### 2.1. Segmentation of Group Delay (GD)

Group delay is a signal processing strategy to determine the syllable section limits in the speech waveform without the information of the transcription [17, 18]. There is no training included also, just the acoustic signs are misused to show up at the syllable limit information. The phrase "carry an oily" pronunciation is demonstrated in figure 3.



**Figure 3** Group delay segments

The syllable level transcription for this phrase is "kaeriyaenoyliy" taking 5 syllables. In any case, the articulation has just 4 syllables "kaeriynoyliy" as proposed by the Group delay based limits. This is once more affirmed by manual listening. This proposes that the transcription is not precise. On the off chance that syllable structure is thought of as a sign to

comment on the waveform, the segment "an" takes VC structure although syllable structure in the waveform is CV. This discrepancy can be used to distinguish inadequately expressed vowels. It is clear to decide the syllable structure from record but to determine the syllable structure from the waveform, a classifier is required. The classifier is explained in the next section.

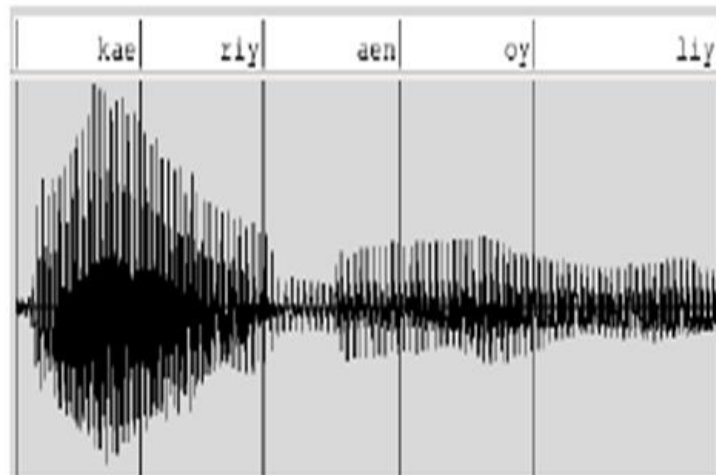


Figure 4 Viterbi alignment

## 2.2. Silence-Vowel-Consonant (SVC) Classifier

The Silence Vowel Consonant classifier is a multilayer perception based classifier inferred by collecting the yield of a multi-layer perception based phoneme recognizer [19]. The multi-layer perceptron-based phoneme recognizer is prepared with 9-outline setting utilizing Modified group delay features [20, 21] in addition to delta coefficients separated from 100 hours of conversational phone discourse translated at phoneme level [22]. There are 45 names, speaking to 17 vowels, 27 consonants and 1 quiet phoneme. The Silence Vowel Consonant classifier classifies the waveform into Silence/Vowel/Consonant at outline level by gathering the separate yields of multi-layer perceptron-based phoneme recognizer. The continuous indistinguishable outputs are then gathered to give different labels in consecutive blocks. This output is smoothed further blending littler measured blocks with their neighboring blocks.

## 2.3. Viterbi Alignment (VA)

A phoneme recognizer is prepared and adjusted for the greatest elocution in the CMU vocabulary [12]. For certain articulations, it is seen that number and locations of the vowels got utilizing group delay Silence Vowel Consonant segmentation and Silence Vowel Consonant classifier don't coordinate that of the record. To connect the transcription with the dissection results, Viterbi alignment is utilized. Each telephone is demonstrated by a 7 state HMM utilizing 3 segment Gaussian mixture density functions. HTK [17] is utilized to prepare the models with the record given in the information base. The prepared models are then adjusted for the best articulation in the CMU word reference [12]. The boundaries of syllable segments are obtained from Viterbi alignment and the syllable portions from Section 2.1 are related. The transcription utilized for alignment and training is acquired from the database, the alignment will yield a limit for each unit in the transcription. For instance, the arrangement of transcription given by Viterbi alignment for the expression "carry an oily" in Figure 3 is appeared in Figure 4. Even though "an" is hardly expressed, a limited number of casings are relegated to "an" in the alignment. The group delay proposes that maybe the syllable is absent. Some Common Mistakes.

## 3. Observations and empirical analysis

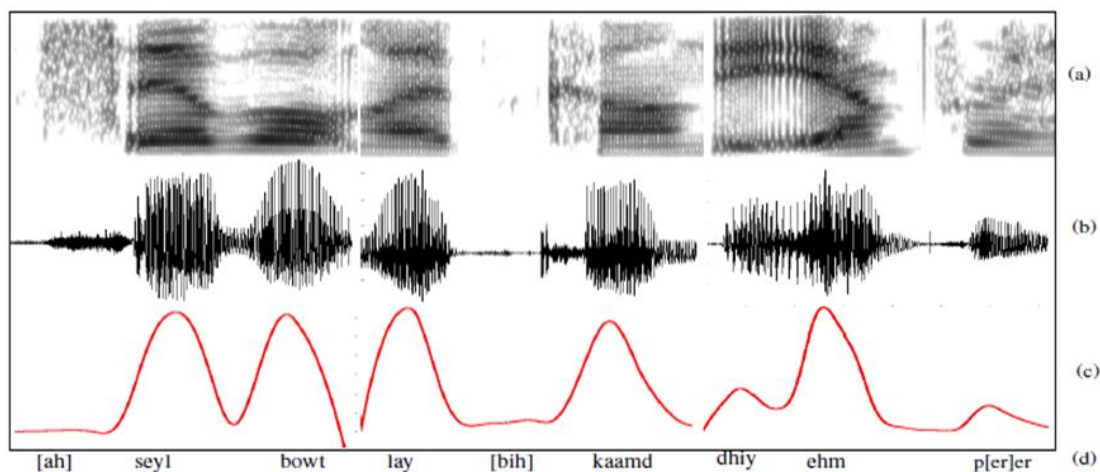
The conversation in Section 2 identifies that individually instrument gives some valuable data about the syllables. We currently attempt to combines the outcomes from all the three tools to decide the right transcription. As referenced before the mismatch is often similar to emphasizing vowels in syllable segments in the expression.

The non-appearance of the vowel is additionally affirmed by Silence Vowel Consonant result. Figure 5 gives graphic delineation of lost vowels in the expression. Intermittently the quantity of group delay portions is greater than the

quantity of character segments in the record. The syllable structure from Silence Vowel Consonant shows the bisyllable example "CVCV" emphasizing the existence of additional vowels for a monosyllable in record. Curiously these words have the accompanying qualities:

- Connecting vowels
- Vowel is followed by "r"

It is seen from the experimental examination that the presence of sequential vowels may occasionally add to two distinct vowels in the expression. For instance, the way to express the word "suit" as given in CMU dictionary [12] seems to be "suwt". For this word the syllable structure is "CVC". Sometimes both the successive vowels "u" and "I" are expressed by the speaker bringing about a bisyllabic (suwiht) word. The syllable structure of "suit" when it brings about a bisyllable articulation is "CVVC". The transition region starting with one vowel then onto the next has a drop in energy ensuing in the "C" locale in SVC. Along these lines both GD and SVC affirm the presence of the extra vowel. In the subsequent case, the "r" following the vowel is regularly articulated as the vowel "er" yielding another vowel. The signals from GD furthermore, SVC again declare this.



**Figure 5** The vowels which are having Poor articulation. (a) Spectrogram (b) Waveform (c) Smoothed energy (d) syllable transcription

#### 4. Automatic annotator

The explanations in Section 3 recommend the advancement of an automatic annotator that emulates speech production. The VA gives the best articulation that matches the expression of a word. This articulation is then syllabified utilizing NIST syllabification programming [23] and the syllable limits are found from VA. An algorithm is currently proposed for revising the transcription:

##### 4.1. Removing poorly highlighted vowel

- The boundaries of syllable from Viterbi alignment and observations from Silence Vowel Consonant and group delay are matched. The missing vowel or a poorly pronounced vowel results in missing margins in group delay.
- The missing vowel is confirmed from Silence Vowel Consonant. In order to eliminate the vowel in function words and word inner bi-phone syllables, the transcription is modified.

##### 4.2. The vowels which are having over stress

- The Articulations of vowels are not recommended in the Articulation dictionary, requires an additional syllable segments in group delay.
- The existence of the additional vowel in group delay is confirmed by Silence Vowel Consonant.
- However the Viterbi alignment does not specify the existence of the additional vowel since, the dictionary does not recommend the same.
- The phonetic transcription of introduced vowel is unidentified so the segments are ignored.

## 5. Experimental results

In order to compare the novel transcription accuracy, a phone level master label file (MLF) is created with all preparation documents. While producing the MLF, the consonants of inadequately highlighted syllables are inserted in proper position. The experiment is accomplished using TIMIT corpus. This new record is then constrained lined up with the models utilized in Section 2.3. The standardized probability of each sentence is then contrasted and the standardized probability of the individual constrained adjusted transcription that comes with the database. Remarkably, the probability of the proposed transcription with erased vowels is high contrasted with the unique transcription. Of the 4620 sentences that are consequently clarified 1402 files (30.35%) have endured missing of vowels in the articulations. In addition to the vowel cancellation, silences are also incorporated in proper places specified by group delay segmentation and SVC classifier. The mandatory deployment of such transcriptions is more likely compared to the original transcription. 1990 sentences which is around 43% in 4620 articulations show improved probability. Noticeably, the increase in probability affirms that the acoustic prompts acquired utilizing signal processing are surely right. The normal relative improve in probability is 0.3%. Regardless of whether such varieties, whenever remembered for the elocution will surely result in better execution is to be considered. This is on the grounds that current discourse acknowledgment frameworks are guided altogether by the language model. Along these lines, unobtrusive improves in probability are probably going to go unseen.

---

## 6. Discussion and Conclusion

This paper tries to apprehend the misalignment among the waveform and the report making use of activates received from system studying and signal processing tools. With spontaneous speech it's far visible that both vowels are lacking or ineffectively expressed or redundant vowels are introduced. The unlucky prominence of vowels is common if the syllable is a function word or a word internal bi-phone syllable. Likewise, phrases/words with connecting vowels or phrases having vowels observed by "r" can provide ascent to new vowels withinside the articulation. The lacking vowel is removed from the transcription, at the same time as the more vowel is disregarded all through probability computation. It is visible that the corrected transcription offers the higher results.

The work is benefited to the human society as a whole, since the diction is an important asset of voice manipulation and grouping. In signing community singers spend more time singing vowels sounds in comparison to the consonance which is why so much importance is placed on them when practicing. Each tongue placement and mouth shape gives the vowels its own characteristic ( known as formants) which identify the vowels to the listeners. Further analysis of the filler stimuli suggests the perceptual advantage of laryngealization (creaky voice) on the previous vowels.

---

## Compliance with ethical standards

### *Acknowledgments*

I wish to acknowledge the support given by Research Centre, Dr. Ambedkar Institute of Technology, Banaglore, Karnataka, India.

### *Disclosure of conflict of interest*

Dr. B. Sivakumar- There are no conflicts of interest in connection with this paper, and the material described is not under publication or consideration for publication elsewhere.

Praveen K B - There are no conflicts of interest in connection with this paper, and the material described is not under publication or consideration for publication elsewhere.

---



## References

- [1] Praveen.K.B, Dr. B.Sivakumar, "Analysis of Resonant Peaks and Pitch for English Vowels of Diversified South Indian English Speakers", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 02-Special Issue, 2019.
- [2] Praveen. K.B, Dr. B. Siva Kumar, "Indian Vowels based Evaluation of First Fundamental Frequency and its Variant Bandwidths" TEST Engineering and Management, January - February 2020 ISSN: 0193 - 4120 Page No. 2166 - 2171.



- [3] Cole, Ronald A., et al. "The contribution of consonants versus vowels to word recognition in fluent speech." 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Vol. 2. IEEE, 1996.
- [4] Kewley-Port, Diane, T. Zachary Burkle, and Jae Hee Lee. "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners." *The Journal of the Acoustical Society of America* 122.4 (2007): 2365-2375.
- [5] Owren, Michael J., and Gina C. Cardillo. "The relative roles of vowels and consonants in discriminating talker identity versus word meaning." *The Journal of the Acoustical Society of America* 119.3 (2006): 1727-1739.
- [6] Parikh, Gaurang, and Philipos C. Loizou. "The influence of noise on vowel and consonant cues." *The Journal of the Acoustical Society of America* 118.6 (2005): 3874-3888.
- [7] Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone, Mark Ordowski, and George R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE transaction on speech and audio processing*, vol. 9, pp. 358–366, 2001.
- [8] Joana Cholin, Niels O. Schiller, and Willem J.M. Levelt, "The preparation of syllables in speech production," *Journal of Memory and Language*, vol. 50, pp.47–61, 2004.
- [9] Steven Greenberg, "Speaking in shorthand - a syllablecentric perspective for understanding pronunciation variation," *Elsevier Speech Communication*, vol. 29, pp. 159–176, 1999.
- [10] Christophe Van Bael, Lou Boves, Henk van den Heuvel, and Helmer Strik, "Automatic phonetic transcription of large speech corpora," *Computer Speech and Language*, March 2007.
- [11] Helmer Strik and Catia Cucchiari, "Modeling pronunciation variation for asr: A survey of the literature," *Elsevier Speech Communication*, vol. 29, pp. 225–246, 1999.
- [12] Kris Demuynck, Tom Laureys, and Steven Gillis, "Automatic generation of phonetic transcriptions for large speech corpora," *Proceedings of Spoken Language Processing*, September 2002.
- [13] "Cmu lexicon," [www.speech.cs.cmu.edu/cgi-bin/cmudict](http://www.speech.cs.cmu.edu/cgi-bin/cmudict).
- [14] Xuedong Huang, Alex Acero, and Hsiao Wuen Hon, *Spoken Language Processing*, Prentice Hall Inc., Upper Saddle River, New Jersey, 2001.
- [15] Raymond W. M. Ng and Keikichi Hirose, "Syllable: A self-contained unit to model pronunciation variation," *ICASSP*, pp. 4457–4460, 2012.
- [16] R. Golda Brunet and Hema A Murthy, "Impact of pronunciation variation in speech recognition," *Proceedings of SPCOM*, July 2012.
- [17] T. Nagarajan and Hema A. Murthy, "Group delay based segmentation of spontaneous speech into syllable-like units," *ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition*, pp. 115–118, Apr 2003.
- [18] T. Nagarajan, V. Kamakshi Prasad, and Hema A. Murthy, "The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation," *Sixth Biennial Conference of Signal Processing and Communications*, July 2001.
- [19] Joel Pinto, G.S.V.S. Sivaram, Mathew Magimai Doss, Hynek Hermansky, and Herve Boulard, "Analysis of mlp based hierarchical phoneme posterior probability estimator," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 225–241, Feb 2011.
- [20] H. A. Murthy and V. R. R. Gadde, "The modified group delay function and its application to phoneme recognition," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 68–71, 2003.
- [21] R. M. Hegde, H. A. Murthy, and V. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 190–202, Jan 2007.
- [22] R. Padmanabhan, *Studies on voice activity detection and feature diversity for speaker recognition*, PhD Thesis, Indian Institute of Technology, Madras, Aug 2012.
- [23] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, Upper Saddle River, New Jersey, 2006.
- [24] Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J.G., Pallett D. S., and Dahlgren N. L., "Timit acoustic phonetic continuous speech corpus," 1993.

### Author's short biography

	<p><b>Dr. B. Sivakumar</b> has got the academic qualification as B.E., M.E., PGDBA, Ph.D. Presently as Professor Dept. of ETE (Govt. Aided), and Graduated from Madurai Kamaraj University in the field of Electronics and Communication. Also obtained his Master degree from PSG College of Technology, Bharathiar University in the field of Applied Electronics. Has been awarded Doctoral Degree in the field of Information &amp; Communication Engineering from Anna University, Chennai. Has got a rich teaching experience of 31 years. Has to his credit 44 International/National in peer-reviewed Journals and 120 Conference Papers. He has already guided 3 Ph.D Scholars and presently guiding 5. He has filed 2 International patent to his credit. He has completed 2 AICTE Research grant projects to the tune of 20 lakhs. Presently carrying out 3 AICTE- AQIS programs. He is an Editor /Reviewer for 4 International Journals.</p>
	<p><b>Praveen K B</b> currently working as Assistant Professor in the department ETE, Dr. Ambedkar Institute of Technology, Bangalore, Karnataka. Presently pursuing doctoral program from PRIST University, Thanjavur, Tamil Nadu. He has got to his credit 6 International/National in peer-reviewed Journals and 5 Conference Papers. Also file 1 patent and is under scrutiny. Have organized several academic programs.</p>