Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 4, April 2017, pg.308 – 315

A Review: Automatic Speech Segmentation

Alaa Ehab Sakran [1], Sherif Mahdy Abdou [23], Salah Eldeen Hamid [4], Mohsen Rashwan [25]
[1] PhD. Program in Computer Science, Sudan University of Science and Technology, Khartoum, Sudan
[2] Research & Development International (RDI®), Giza, Egypt
[3] Department of IT, Faculty of Computers and Information, Cairo University, Giza, Egypt
[4] Department of Engineering and Applied Sciences, Umm Al-Qura University, Mecca, Saudi Arabia
[5] Department of Electronics and Communication Engineering, Cairo University, Giza, Egypt
alaa238.ehab@gmail.com, salahamed@gmail.com, {sabdou, mrashwan} @rdi-eg.com

I. INTRODUCTION

Automated segmentation of speech signals has been under research for over 30 years. Many speech processing systems require segmentation of Speech waveform into principal acoustic units. Segmentation is a process of breaking down a speech signal into smaller units. Segmentation is the very primary step in any voiced activated systems like speech recognition systems and training of speech synthesis systems. Speech segmentation is performed utilizing Wavelet, Fuzzy methods, Artificial Neural Networks and Hidden Markov Model. [1] [2]

In this research a review of basics of speech segmentation problem and state-of-the art solutions will be investigated. In next section main characteristics of speech signal will be discussed. In section 3 main aspects of speech segmentation problem will be shown. Trends in choosing segmentation units are discussed in section 4. Types of speech segmentation are investigated in section 5. Section 6 deals with main speech features that are used for segmentation. Different approaches used in assessment of automatic segmentation process are discussed in section 6. Main methods used in segmentation of speech are investigated in section 7.

II. GENERAL CHARACTERISTICS OF SPEECH

A continuous speech signal consists of two main parts: one carries the speech information, and the second carries the silent or noise sections that are in between the utterances, without any verbal information.

The verbal (informative) part of speech can be further divided into two main types: Voiced and Unvoiced speech. While humans speak all air paths through larynx, voiced sounds are generated when vocal cords are semi closed while when they are opened unvoiced speech is generated.

In voiced speech is a relatively slowly changing periodic signal with case frequency of vibration of vocal cords called pitch. Male's contribution of pitch is commonly in the range between 50Hz to 250Hz while female's contribution of pitch lies between 120Hz and 500Hz.

Unvoiced speech sounds are produced by air passed directly through vocal tract formations. Unlike voiced speech, unvoiced speech does not exhibit periodicity, and is characterized by a noise-like signal.

The speech production process involves producing speech utterances which are groups of successive voiced and/or unvoiced sounds. These sounds usually are very smoothly connected.

Speech utterances are separated by silence regions. There is no excitation supplied to the vocal tract during silence regions and hence there is no speech output. None-the-less, silence is considered an integral part of speech signal [3]. Acoustic signals during silence regions are mainly background noise and speech irrelevant mouth sounds.

III. SEGMENTATION

Speech segmentation can be defined as the process of finding the limits (with specific characteristic) in natural spoken language between words, syllables or phonemes. [4], [5]

The main objective of Speech segmentation is to serve other speech analysis problems such as speech synthesis, data training for speech recognizers, or to fabricate and label prosodic databases. Therefore, it can be viewed as a vital sub-issue for various fields in speech analysis and research. [6], [7]

The traditional approach handling this issue is by manual segmentation of speech, which is generally performed by specialized phoneticians. However, this method is based on listening and visual judgment on required boundaries which makes it inconsistent and time consuming. [8], [9]

Another method which is considered very convenient is an automatic segmentation. The speech can be automatically segmented into sub word units which are defined acoustically. [10]

In Automatic speech Recognition ASR systems, segmentation can be performed:

(i) At the system training stage, when segmentation is applied to the training set recordings.

(ii) At the recognition stage. [5]

IV. SEGMENTATION UNITS

The Speech recognition and synthesis systems need a speech signal to be segmented into some basic units like Words, Phonemes, or syllables. Depending on the extent or size of vocabulary the decision of representative units is made. Word is the most natural unit of segmentation. It's not suitable to use words as the units for segmentation because of the absence of generalization and more memory space consumption. [10]

Phonemes are the smallest segmental unit of sound employed to form meaning. The same phoneme in various words has distinctive significance. There is an over generalization of phonemes. So the blend of phoneme and words gives rise to next level basic unit of speech called as syllables. Syllable like units are defined by rules, a syllable must have a vowel called its nucleus. [11]

The realization of a phoneme is strongly influenced by its adjacent phonemes. Phonemes are highly context dependent. Hence, the acoustic variability of basic phonetic units due to context is extremely large and is not well understood in many languages. [12][10]

V. TYPES OF AUTOMATIC SPEECH SEGMENTATION

Automatic speech segmentation strategies can be grouped in various perspectives, however one very common classification is the division to blind and aided segmentation algorithms.

1) Blind segmentation: The term blind segmentation refers to methods where there is no pre-existing or external knowledge regarding linguistic properties, such as orthography or full phonetic annotation of the signal to be segmented. Blind segmentation is applied in various applications, for example speaker verification systems, speech recognition systems, language identification systems, and speech corpus segmentation and labeling [13].

Solutions to this problem comprise of algorithms which do not require any background knowledge about the phonetic content and are based predominantly on statistical signal analysis [26]. [5]

Due to the lack of external or top-down information, the first phase of blind segmentation relies completely on the acoustical features present in the signal. The second phase or bottom-up processing is normally built on a front-end parameterization of the speech signal, often using MFCC, LP-coefficients, or pure FFT spectrum. [14]

2) Aided segmentation: Aided segmentation Algorithms use some sort of external linguistic knowledge of the speech stream to segment it into relating segments of the wanted type. An orthographic or phonetic transcription is used as a parallel contribution with the speech, or training the algorithm with such data. [15] These algorithm types are computationally consuming. This group includes algorithms using recognition with Hidden Markov Models (HMMs) [21], [22], [23], Dynamic Time Warping (DTW) [24] or Artificial Neural Networks (ANNs) [25]. The algorithms of this group are employed only in the ASR system training stage.

One of the most common methods in ASR for utilizing phonetic annotations is with HMM- based systems [16]. HMM-based algorithms have dominated most speech recognition applications since the 1980s because of their high performance in recognition and relatively small computational complexity in the field of speech recognition [17][18][19][20].

VI. TYPES OF FEATURES ARE USED FOR SEGMENTING

Two types of signal features are used for segmenting speech signal: time-domain features and frequency-domain features.

A. Time-Domain Signal Features

Time-domain features are widely used for speech segment extractions. These features are useful when it is needed to have algorithm with simple implementation and efficient calculation.

1) Short-Time Signal Energy: Short-time energy is the dominant and most natural feature that has been used. Physically, energy is a measure of how much signal there is at any one time. Energy is used to discover voiced sound, which have higher energy than silence/unvoiced sound in a continuous speech. The energy of a signal is typically calculated on a short-time basically by windowing the signal at a particular time, squaring the samples and taking the average. [27], [28]
The square root of this result is an engineering quantity, known as the Root Mean Square (RMS) value. The short-time energy function of a speech frame with length N is defined as:

$$E_n = \frac{1}{N}\sum_{m=-\infty}^{\infty}[x(n-m)w(m)]^2 \qquad (1)$$

The short-term root mean square (RMS) energy of this frame is given by:

$$E_{n(RMS)} = \sqrt{\frac{1}{N}\sum_{m=-\infty}^{\infty}[x(n-m)w(m)]^2} \qquad (2)$$

Where x(n) is the discrete-time audio signal and w(n) is rectangle window function. [1], [29], [30], [31]

$$w(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & otherwise \end{cases} \qquad (3)$$

2) Short-Time Average Zero-Crossing Rate: The average zero-crossing rate refers to the number of times speech samples change algebraic sign in a given frame. [27]
The rate at which zero crossings occur is a simple measure of the frequency content of a signal. It is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. [32] Unvoiced speech components normally have much higher ZCR values than voiced ones.

The short-time average zero-crossing rate is defined as:

$$Z_n = \tfrac{1}{2}\textstyle\sum_{m=-\infty}^{\infty}|sgn[x(n-m)] - sgn[x(n-m-1)]|w(m) \qquad (4)$$

Where,

$$sgn[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \qquad (5)$$

And w (n) is a rectangle window of length N, given in equation (3). [1], [30]

B. Frequency-Domain Signal Features

The most information of speech is amassed in 250Hz-6800Hz frequency range. In order to extract frequency-domain features, discrete Fourier transform can be used. The Fourier representation of a signal demonstrates the spectral composition of the signal. [33]

1) Spectral Centroid: The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of gravity of the spectrum high values corresponding to brighter sounds. [34] [35]
The spectral centroid, SCi , of the i-th frame is defined as the center of gravity of its spectrum and it is given by the following equation:

$$SC_i = \frac{\sum_{n=0}^{N-1} f_i(n) x_i(n)}{\sum_{n=0}^{N-1} x_i(n)} \qquad (6)$$

Where x(n) represent the weighted frequency value, or magnitude, of bin number n, and f(n) represents the center frequency of that bin in DFT spectrum. The DFT is given by the following equation and can be computed efficiently using a fast Fourier transform (FFT) algorithm. [33]

$$X_k = \sum_{n=0}^{N-1} x_n . e^{-j2\pi k \frac{n}{N}} \quad , \quad k = 0, \ldots, N-1 \quad (7)$$

Here, X_k is the DFT coefficients of i-th short-term frame with length N. [29], [1], [30]

2) Spectral flux: Spectral flux is a measure of how quickly the power spectrum of a signal is changing calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame, also known as the Euclidean distance between the two normalized spectra. The spectral flux can be used to determine the timbre of an audio signal, or in onset detection [36] among other things. The equation of Spectral Flux, SF_i is given by:

$$SF_i = \sum_{k=1}^{N/2} (|X_i(k)| - |X_{i-1}(k)|)^2 \qquad (8)$$

Here, Xi(k) is the DFT k-th coefficient of i-th short-term frame with length N, given in equation (7). [1], [30]

VII. AUTOMATIC SPEECH SEGMENTATION METHODS

Some of the various methods involved in automatic segmentation are described below.

A. Wavelet Method

B. Zioko, S. Manandhar, Richard C. Wilson and M. Zioko has discussed the technique of the wavelet method to recognize the beginnings and ends of the phonemes based on (DWT) analysis and with the aid of spectral analysis, it happened to be an extremely effective technique. [37]

S. Ratsameewichai, N. Theera-Umpon, J. Vilasdechanon, S. Uatrongjit, and K. Likit- Anurucks have divided the speech into low and high frequency components by breaking down the wavelet. Then they used the energy contour to identify the boundaries of the phoneme. They have performed the experiment using 1,000 syllables data recorded from 10 speakers. The accuracy rates are 96.0, 89.9, 92.7 and 98.9% for initial consonant, vowel, final consonant and silence, respectively. [35]

M. F. Tolba, M. E. Gadallah, T. Nazmy and A. A. Abdelhamid has presented a technique without any linguistic information and based on wavelet transform and spectral analysis focused on searching the transient between consonant and vowel parts. He performed the experiment using a set of 20 Arabic words of which each of it recorded 6 times.

The accuracy rate obtained was about 88.3 % for Consonant-vowel segmentation. [38]

B. Artificial Neural Networks

J. Kamarauskas has presented a technique using perceptron and back propagation artificial neural network to recognize distinctive phonemes, by utilizing various features of the speech signal used in speech or speaker recognition (linear prediction coding, Cepstral coefficients, and coefficients of the Fourier transform). He also determined that artificial neural networks can also be used in setting beginning and end points of the word and separate voiced and non-voiced phonemes.

From his experiments he noted that the error rate of recognition of the back propagation artificial neural network is lower than that of perceptron, but its training is much longer [39]

Y. Suh and Y. Lee have discussed the technique of implemented phoneme segmentation method using MLP (multi-layer perceptron). The segmenter consisted of three elements preprocessor, MLP based phoneme segmenter, and post processor. The pre-processor utilized a sequence of 44 order feature parameters for each frame of speech, based on the acoustic-phonetic knowledge. The MLP has one hidden layer and an output layer. The feature parameters for four consecutive inter-frame features (176 parameters) are served as input data. The output value decides whether the current frame is a phoneme boundary or not. In post processing, they have decided the positions of phoneme boundaries using the output of the MLP.

They have obtained 84 % for 5 msec accuracy and 87 % for 15 msec accuracy. When they decreased the threshold by 0.4, they obtained 5 msec accuracy of 92 % with insertion rate of 3.4 % for the insertions that are more than 15 msec apart from phoneme boundaries (The insertion rate is the ratio of the number of frames incorrectly decided as boundary to that of total non-boundary frames). [40]

C. Blocking Black Area Method

Md. Mijanur Rahman, F. Khatun and Md. Al- Amin Bhuiyan offered a method for automatic speech segmentation named blocking black area to block the voiced regions of the continuous speech signal to distinguish voiced from unvoiced (silence) by utilizing Ostu's method for dynamic threshold, The edges of the block are used as word boundaries in the continuous speech. To test the performance of the system, the database contains 500 Bangla sentences with 3280 words. All the algorithms and methods used in this research are implemented in MATLAB and the proposed system has achieved the average segmentation accuracy of 90.58. [2]

D. Short Term Energy

E. A. Kaur and E. T. Singh have used Short Term Energy of speech method for segmentation of speech into syllables. Technique has been implemented in Matlab 7.8. Various speech signals in Punjabi have been recorded and segmented. Proposed method has been implemented and analyzed for different Punjabi speech signals. [10]

S. Hossain, N. Nahid, N. Nuzhat Khan , D. Gomes, S. Mohammad Mugab has a technique of Short Term Energy of Speech Signal. The overall accuracy of speech samples was about 85%. [41]

E. Hybrid Speech Segmentation Algorithm

M. Kalamani ,S. Valarmathy and S. Anitha have discussed the technique of hybrid segmentation method proposed for Automatic Speech Recognition. The segmentation methods are based on time domain features and frequency domain features. The time domain features are short time energy, short time zero crossing rate. The frequency domain features are spectral centroid and spectral flux. The features are extracted then a simple threshold methodology is used to detect the boundaries of the word. Hence the segmentation is characterized to breakdown continuous speech into a sequence of words or sub words. This method achieved segmentation accuracy of 98.33% and error rate is 1.67%. [30]

F. Word Chopper Technique

N. Sharma and P. Singh have proposed the technique of segmentation of speech into syllables using Word Chopper. They proposed a new approach which has three stages comprising of feature extraction, Rule Matching and segmentation. [42]

G. Hidden Markov Model

P. Bansal, A. Pradhan, A. Goyal, A. Sharma, M. Arora has used for implementation for phonetic segmentation and analysis of speech at phonetic level. The accuracy rate obtained was about 78.14%. An observation by the researcher was more the states better the alignment and precision obtained during modeling. [43]

J. Dines, S. Sridharan and M. Moody have explained the aspects of their automatic speech segmentation system that they have used in conjunction with their speech synthesis research. It was based on a Hidden Markov Model phone recognizer by using training strategies optimized for the segmentation task. They illustrated their phone recognizer design and identified the differences in paradigms estimation for speech segmentation. System evaluation demonstrates the ability of their system to provide high reliability speech segmentation. [44]

A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang and M. Liberman have presented techniques for boosting the accuracy of automatic phonetic segmentation based on HMM acoustic-phonetic models. They have shown improved test results by using more powerful statistical models for boundary correction that are conditioned on phonetic context and duration features. They concluded that combining multiple acoustic front-ends gave an additional gain in accuracy, and that conditioning the combiner on phonetic context and side information helps with results, which reduced segmentation errors on the TIMIT corpus by almost one half, from 93.9% to 96.8% boundary accuracy with a 20-ms tolerance. [45]

VIII. CONCLUSION

Key papers of techniques and methodologies have been read. Among the techniques investigated the short term energy technique which suffered from the problem of thresholding. The Word Chopper based segmentation cannot be used for segmentation of all words. In the neural network the error rate of recognition of the back propagation artificial neural network is lower than that of perceptron, but its training is much longer. One of the efficient methods for segmentation of speech is Discrete Wavelet Transform (DWT) because this method uses the frequency and time simultaneously. The most accurate and efficient methodology researched was hidden Markov model as the results which were obtained was approximately same as obtained using manual segmentation. [43]

REFERENCES

[1] M. M. Rahman and M. Bhuiyan, “Continuous bangla speech segmentation using short-term speech features extraction ap- proaches,” International Journal of Advanced Computer Sci- ences and Applications, vol. 3, no. 11, p. 485, 2012.

[2] M. M. Rahman, F. Khatun, and M. A.-A. Bhuiyan, “Blocking black area method for speech segmentation,” Editorial Preface, vol. 4, no. 2, 2015.

[3] M. Nilsson and M. Ejnarsson, “Speech recognition using hidden markov model,” 2002.

[4] M. A. Al-Manie, M. I. Alkanhal, M. M. Al-Ghamdi, N. Mas- torakis, A. Croitoru, V. Balas, E. Son, and V. Mladenov, “Auto- matic speech segmentation using the arabic phonetic database,” in WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering, no. 10. World Scientific and Engineering Academy and Society, 2009.

[5] R. Makowski and R. Hossa, “Automatic speech signal segmentation based on the innovation adaptive filter,” International Journal of Applied Mathematics and Computer Science, vol. 24, no. 2, pp. 259–270, 2014.

[6] A. Cherif, L. Bouafif, and T. Dabbabi, “Pitch detection and for- mant analysis of arabic speech processing,” Applied Acoustics, vol. 62, no. 10, pp. 1129–1140, 2001.

[7] M. Sharma and R. Mammone, “Subword-based text-dependent speaker verification system with user-selectable passwords,” in Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 1. IEEE, 1996, pp. 93–96.

[8] J. P. van Hemert, “Automatic segmentation of speech,” IEEE Transactions on Signal Processing, vol. 39, no. 4, pp. 1008–1012, 1991.

[9] Y.-S. Lee, K. Papineni, S. Roukos, O. Emam, and H. Hassan, “Language model based arabic word segmentation,” in Proceed- ings of the 41st Annual Meeting on Association for Computa- tional Linguistics-Volume 1. Association for Computational Linguistics, 2003, pp. 399–406.

[10] E. A. Kaur and E. T. Singh, “Segmentation of continuous punjabi speech signal into syllables,” in Proceedings of the World Congress on Engineering and Computer Science, vol. 1. Citeseer, 2010, pp. 20–22.

[11] T. Nagarajan, H. A. Murthy, and R. M. Hegde, “Segmentation of speech into syllable-like units,” Energy, vol. 1, no. 1.5, p. 2, 2003.

[12] R. Thangarajan and A. Natarajan, “Syllable based continuous speech recognition for tamil,” South Asian language review, vol. 18, no. 1, pp. 72–85, 2008.

[13] M. Sharma and R. Mammone, “” blind” speech segmentation: Automatic segmentation of speech without linguistic knowl- edge,” 1996.

[14] A. SaiJayram, V. Ramasubramanian, and T. Sreenivas, “Robust parameters for automatic segmentation of speech,” in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE Interna- tional Conference on, vol. 1. IEEE, 2002, pp. I–513.

[15] F. Schiel, “Automatic phonetic transcription of non-prompted speech,” 1999.

[16] K. Knill and S. Young, “Hidden markov models in speech and language processing,” in Corpus-based methods in language and speech processing. Springer, 1997, pp. 27–68.

[17] B.-H. Juang and L. R. Rabiner, “Automatic speech recognition– a brief history of the technology development,” Georgia Insti- tute of Technology. Atlanta Rutgers University and the Univer- sity of California. Santa Barbara, vol. 1, p. 67, 2005.

[18] N. Chowdhury, M. A. Sattar, and A. K. Bishwas, “Separating words from continuous bangla speech,” Global Journal of Computer Science and Technology, vol. 4, pp. 172–175, 2009.

[19] C.-T. Hsieh, “Segmentation of continuous speech into phonemic units,” IEICS, pp. 420–424, 1991.

[20] M. M. Rahman, M. F. Khan, and M. A. Moni, “Speech recognition front-end for segmenting and clustering continuous bangla speech,” Daffodil International University Journal of Science and Technology, vol. 5, no. 1, pp. 67–72, 2010.

[21] H. M. M. Tanqueiro, “Utilització didiomes.”

[22] D. T. Toledano, L. A. H. Gómez, and L. V. Grande, “Automatic phonetic segmentation,” IEEE transactions on speech and audio processing, vol. 11, no. 6, pp. 617–625, 2003.

[23] I. Mporas, T. Ganchev, and N. Fakotakis, “A hybrid architecture for automatic segmentation of speech waveforms,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008, pp. 4457–4460.

[24] J. A. Gómez and M. Calvo, “Improvements on automatic speech segmentation at the phonetic level,” in Iberoamerican Congress on Pattern Recognition. Springer, 2011, pp. 557–564.

[25] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, “High-accuracy phone recognition by combining high-performance lattice gen- eration and knowledge based rescoring,” in 2007 IEEE Interna- tional Conference on Acoustics, Speech and Signal Processing- ICASSP’07, vol. 4. IEEE, 2007, pp. IV–869.

[26] O. Scharenborg, V. Wan, and M. Ernestus, “Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries,” The Journal of the Acoustical Society of America, vol. 127, no. 2, pp. 1084–1095, 2010.

[27] T. Zhang and C.-C. Kuo, “Hierarchical classification of audio data for archiving and retrieving,” in Acoustics, Speech, and Sig- nal Processing, 1999. Proceedings., 1999 IEEE International Conference on, vol. 6. IEEE, 1999, pp. 3001–3004.

[28] G. Hemakumar and P. Punitha, “Automatic segmentation of kannada speech signal into syllables and sub-words: noised and noiseless signals,” International Journal of Scientific & Engineering Research, vol. 5, no. 1, pp. 1707–1711, 2014.

[29] I. Khaing and L. KZin, “Automatic speech segmentation for myanmar language,” 2014.

[30] M. Kalamani, S. Valarmathy, and S. Anitha, “Hybrid speech segmentation algorithm for continuous speech recognition.”

[31] J.-J. Chen, , and L.-S. Lee, “Automatic segmentation techniques for mandarin speech recognition,” in International Computer Symposium, Tamkang University, Taipei (1988.12), 1988.

[32] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” Bell System Technical Journal, vol. 54, no. 2, pp. 297–315, 1975.

[33] R. Niederjohn and J. Grotelueschen, “The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp.277–282, 1976.

[34] T. Giannakopoulos, “Study and application of acoustic infor- mation for the detection of harmful content, and fusion with visual information,” Department of Informatics and Telecom- munications, vol. PhD. University of Athens, Greece, 2009.

[35] S. Ratsameewichai, N. Theera-Umpon, J. Vilasdechanon, S. Ua- trongjit, and K. Likit-Anurucks, “Thai phoneme segmentation using dual-band energy contour,” ITC-CSCC: 2002 Proceed- ings, pp. 111–113, 2002.

[36] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, “A tutorial on onset detection in music signals,” IEEE Transactions on speech and audio processing, vol. 13, no. 5, pp. 1035–1047, 2005.

[37] B. Ziółko, S. Manandhar, R. C. Wilson, and M. Ziółko, “Wavelet method of speech segmentation,” in Signal Processing Conference, 2006 14th European. IEEE, 2006, pp. 1–5.

[38] M. Tolba, T. Nazmy, A. Abdelhamid, and M. Gadallah, “A novel method for arabic consonant/vowel segmentation using wavelet transform,” International Journal on Intelligent Coop- erative Information Systems, IJICIS, vol. 5, no. 1, pp. 353–364, 2005.

[39] J. Kamarauskas, “Automatic segmetation of phonemes using ar- tificial neural networks,” Elektronika ir Elektrotechnika, vol. 72, no. 8, pp. 39–42, 2015.

[40] Y. Suh and Y. Lee, “Phoneme segmentation of continuous speech using multi-layer perceptron,” in Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, vol. 3. IEEE, 1996, pp. 1297–1300.

[41] A. Hossain, N. Nahid, N. N. Khan, D. C. Gomes, and S. M. Mugab, “Automatic silence/unvoiced/voiced classification of bangla velar phonemes: New approach,” 8th ICCIT, Dhaka, 2005.

[42] S. Nishi and S. Parminder, “Automatic segmentation of wave file,” Int J of Comput Sci Commun, vol. 1, no. 2, pp. 267–270, 2010.

[43] P. Bansal, A. Pradhan, A. Goyal, A. Sharma, and M. Arora, “Speech synthesis-automatic segmentation,” International Journal of Computer Applications, vol. 98, no. 4, 2014.

[44] J. Dines, S. Sridharan, and M. Moody, “Automatic speech segmentation with hmm,” in Proceedings of the 9th Australian Conference on Speech Science and Technology, 2002.

[45] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liber- man, “Highly accurate phonetic segmentation using boundary correction models and system fusion,” in 2014 IEEE Interna- tional Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 5552–5556.