Research Journal of
Applied Science
MEDWELL PUBLICATIONS

A Survey: Text Independent Automatic Speech Segmentation Techniques

Ihsan Al-Hassani, Oumayma Al-Dakkak and Abdlnaser Assami
Department of Telecommunication, Higher Institute for Applied Science and Technology (HIAST), Damascus, Syria

Key words: ASR, TTS, phonetic segmentation, text-independent, fusion, performance metrics, Query-by-Example (QbyE), modeling, HMM

Corresponding Author:
Ihsan Al-Hassani
Department of Telecommunication, Higher Institute for Applied Science and Technology (HIAST), Damascus, Syria

Page No.: 75-84
Volume: 16, Issue 2, 2021
ISSN: 1815-932x
Research Journal of Applied Sciences

Abstract: Speech segmentation techniques have made important advances in the past decades and are still an active area of research and development. It is a process of breaking down a speech signal into smaller units such as phonemes. Speech segmentation is decisive for many acoustic systems essentially Automatic Speech Recognition (ASR). Phonetic segmentation techniques are divided into two major categories: Text-Dependent (TD) and Text-Independent (TI). In the text-dependent segmentation techniques, the phonetic annotation of the speech signal is already known and we only need to find the boundaries of each phoneme segment. While in the Text-Independent (TI) techniques no annotation is available, thus, the segmentation relies solely on the acoustic information contained in the speech signal. In this study, we present a thorough survey of the different algorithms and techniques proposed so far for solving the problem of text-independent phonetic segmentation.

INTRODUCTION

The phonetic segmentation technique is about identifying the starting and ending boundaries of each phoneme segment in continuous speech. It is an important technique in many areas of speech processing[1, 2]. It can benefit segment-based speech recognition systems[2] which integrate the dynamics of speech better than frame-based ones. Phoneme segmentation is also crucial for creating phoneme databases used in text to speech (TTS) systems[3-5], to transcribe speech corpus used in training HMMs (Hidden Markov Models) in ASR systems. Phonetic segmentation is also used in building a Query-by-Example (QbyE) Spoken Term Detection (STD) application which is relatively a new application drawing increasing attention in recent years[6]. Knowledge of phoneme boundaries is also necessary in some cases of health-related research on human speech processing[6] such as diagnostic marker for Childhood Apraxia of Speech (CAS) and Alzheimer's disease[7]. Phonetic segmentation and annotation can be done either automatically or manually by expert phoneticians[1]. The main difficulty of this task is its subjectivity because of the lack of distinct physiological or acoustic events that signal a phoneme boundary in some cases. In continuous speech, phoneme boundaries are sometimes difficult to locate due to glottalization extremely reduced vowels, or gradual decrease in energy before a pause[7]. As a result, there is no "correct" answer to the phoneme segmentation problem. Instead a measure of the agreement between two alignments is take place, such as the agreement between two humans or the agreement between human and machine[7]. Though manual segmentation is the most adequate[8] way for phonetic transcription but it suffers from being very tedious and time consuming task (it is reported that manual alignment takes between 11 and

30 sec per phoneme[7]), especially in the case of large speech corpora and spontaneous speech. In addition manual segmentation suffers from labeler subjectivity and may not be able to maintain labeling consistency[9]. These difficulties stimulate the development of algorithms for automatic phonetic segmentation of continuous speech waveforms. Automatic speech segmentation techniques are divided into two major categories: Text-Dependent (TD) and Text-Independent (TI) segmentation[10, 11]. Most text dependent segmentation techniques (It also called explicit because we know explicitly the phonetic annotation a priori. Sometimes it is also called linguistically constrained segmentation methods)are based on HMM with forced alignment Viterbi algorithm[10, 12]. These methods suffer many shortages: To provide a good performance, it needs accurate phoneme models that incorporate pronunciation variants and other phonetic phenomena like elision, dialectal variation, cross-word assimilation, de-gemination. Hesitations, false-starts and other dysfluencies which are very common in spontaneous speech are other sources of problems[10]. The corresponding text that matches the speech waveform is not available in many cases, including real-time phoneme-based speech recognition, accent conversion system, real-time translation system and computer aided language learning system[13]. Imposing linguistic constraints to the segmentation algorithms make these algorithms restricted to the database used for training[11]. In the case of foreign or accented speech processing, there can exist a large mismatch between utterances and native acoustic models which degrades the performance of the HMM-based segmentation[14]. All these issues can be handled more efficiently by Text-Independent (TI) segmentation methods (also called implicit), that do not corporate any prior information about the corresponding phonetic or word transcription of the speech waveform to be segmented.

TI methods can be classified into two broad categories: model-based methods and model-free methods. Models based methods incorporate an acoustic modeling stage that can help in discriminating borders from non-borders (phonemes) segments. After learning the acoustic model, segmentation is done through binary classification. The acoustic modeling can be done either using supervised techniques that need a manually segmented training dataset, or unsupervised techniques that can learn the model without the need for any training dataset (We should note here that in some papers[15] blind phonetic segmentation methods are also called blind in the sense that they do not a manually segmented dataset for training. Here, in our classification, we emphasis the difference between unsupervised methods that include a training stage with a non-labeled dataset and blind methods that do not incorporate any training stage). Recently self-supervised learning technique was proposed[15]. In Self Supervised Learning (SSL) methods, the unlabeled input is used to define an auxiliary task that can generate labeled pseudo training data. This can then be used later to train the model using supervised techniques[15]. In model-free methods (also called blind) the segmentation is done by trying to identify phoneme boundaries as spectral changes in the speech signal, directly considering the speech spectrum coupled with several spectral distortion and metric measures, without considering any modeling stage[14, 10, 16]. Figure 1 depicts the classification of different phonetic segmentation systems.

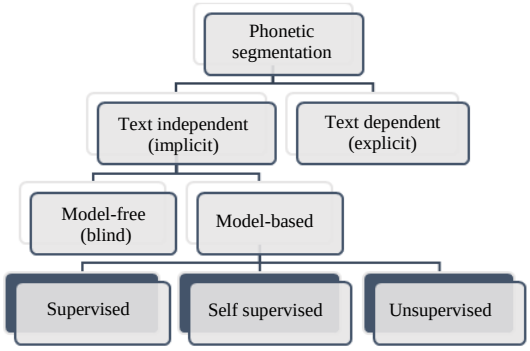

Fig. 1: Classification of phonetic segmentation techniques

In this research, we present an inclusive survey of the Text-Independent (TI) phonetic segmentation techniques.

TI SEGMENTATION ALGORITHMS PERFORMANCE METRICS

In the case of text independent segmentation techniques, the number of discovered segments might differ from the number of segments produced by manual segmentation. TI segmentation can be viewed as a boundary detection problem. Thus, there are two types of errors, Type I error and Type II error.

Type I error is a false alarm or an insertion error. It happens when there are boundaries detected by the algorithm that do not have corresponding boundaries in the reference signal. Insertion errors are expressed by FAR (False Alarm Rate)[17] and OS (Over Segmentation)[18] and are calculated as follows:

$$FAR = \frac{\text{Inserted boudaries}}{\text{All non boudary points}} = \frac{IB}{AP\text{-}ACB}100\% \quad (1)$$

$$OS = \left(\frac{ADB}{ACB}\text{-}1\right)100\% = \frac{ADB\text{-}ACB}{ACB}100\% \quad (2)$$

Where:

ADB = The number of All Detected Boundaries (true and false) by the algorithm

ACB = The Actual Number of Boundaries

IB = The number of Inserted Boundaries (false detection)
AP = All overall points (i.e., the total number of points)

Type II error is a deletion or miss; it happens when there is a boundary marked in the reference, not detected by the algorithm (miss). This error is expressed by the MDR (Miss Detection Rate):

$$MDR=\frac{MD}{ACB}100\% \quad (3)$$

where, MD is the number of undetected boundaries (missed). A high value of FAR means over-segmentation of the speech signal happened. While a high value of MDR means that the algorithm does not segment the audio signal properly. We can see through Eq. 2 and 4 that a higher detection performance (lower MDR) comes at expense of a higher FAR[17].

TI segmentation algorithms can be assessed by two other metrics; hit rate and precision (PRC). When a detected boundaries match corresponding boundary in the reference signal, this is called a hit rate (also called Recall RCL). It can be calculated as follows:

$$\text{Hit rate} = RCL = \frac{CDB}{ATB}100\% \quad (4)$$

where, CDB is the number of Correctly Detected Boundaries. A Precision (PRC) metric can be calculated as follows:

$$PRC=\frac{CDB}{CDB+IB}=\frac{CDB}{ADB}100\% \quad (5)$$

The overall objective effectiveness of the segmentation algorithm can be evaluated by the F1-measure. It is calculated according to the following formula:

$$F1=\frac{2PRC.RCL}{PRC+RCL} \quad (6)$$

F1-measure is the harmonic mean of recall and precision, that is used for assessing classification and prediction algorithms. F1-measure takes value in the unit interval between [0 1] where the score closer to 1 is better. A system with high recalls but low precision returns many results but most of its predicted labels are incorrect. A system with high precision but low recall is just the opposite, returning very few results but most of its predicted labels are correct. An ideal system with high precision and high recall will return many results with all results labeled correctly[17]. F1-score is not suitable for segmentation, optimizing the operation of a speech segmentation algorithm is often a tradeoff between hit-rate and over-segmentation (or inversely, false-alarm rate and miss-rate)[18]. F1-score (7) is one possible way to describe overall performance of an algorithm with a single value. However, F1-score is prone to stochastic hit-rate increases due to the over-segmentation issue[18]. In some cases, over-segmentation may result in very high recall rate causing high F1-score, though the precision might be relatively low.

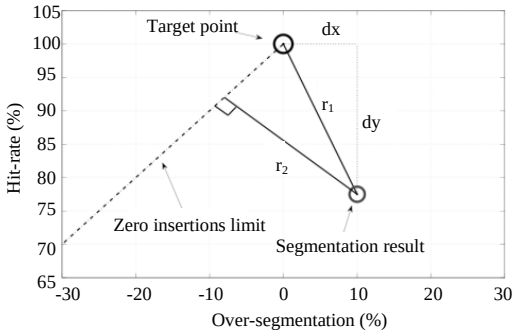

Fig. 2: Calculating R-measure[18]

Rasanen *et al*.[18] proposed a new metric to describe performance using a single value that properly penalizeover-segmentation. The optimal goal of segmentation is to achieve a hit-rate of 100% and an over-segmentation of 0%, this is called the Target Point (TP). The basis of the new metric is the algorithm's distance from TP and not the (hit-rate) gain achieved by over-segmentation.

On the segmentation performance plane illustrated in Fig. 2, a distance r_1 is derived (Eq. 7) and a distance r_2 is measured (Eq. 8), to appreciate the value of under-segmentation compared to over-segmentation in the algorithm (i.e., less false positives). The distances r_1 and r_2 are then added together and normalized to have a maximum value of 1 at the target-point (Eq. 9). This new distance measure, referred to as the R-value, decreases as the distance to the target grows, similarly as F1-score does but it makes more emphasis on over-segmentation by arguing that better hit rates might be achieved by simply adding random boundaries without any algorithmic improvement. This measure evaluates how close one is to the ideal segmentation R = 1. In the literature review we present, we found three studies used this measure in evaluation[19, 20]. Figure 3 depicts the different metrics used for evaluating TI phonetic segmentation algorithms:

$$r_1=\sqrt{(100-HR)^2+(OS)^2} \quad (7)$$

$$r_2=\frac{-OS+HR-100}{\sqrt{2}} \quad (8)$$

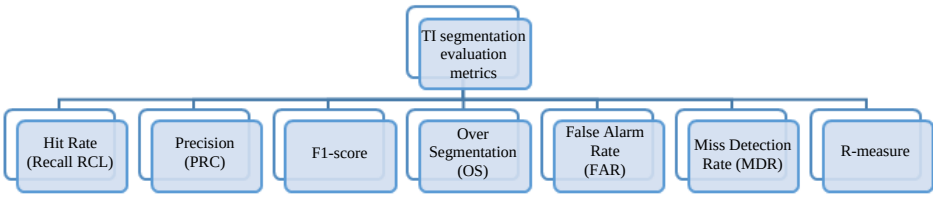

Fig. 3: Different metrics used for evaluating TI phonetic segmentation techniques

$$R=1-\frac{abs(r_1)+abs(r_2)}{200} \quad (9)$$

TEXT INDEPENDENT (TI) PHONETIC SEGMENTATION TECHNIQUES

In contrast with TD segmentation methods, TI methods do not need any phonetic annotation data for the speech signal to be segmented. Instead, they are generally based on sets of rules derived from encoding human knowledge to segment speech, like acoustic rate of change, or other spectral variation metrics[21-24]. Such methods are called blind or model-free because they do not use any modeling stage. Recently several studies proposed using different supervised and unsupervised machine learning techniques to build an acoustic model that can be used in the phoneme segmentation task[25, 26]. These methods are called model-based methods.

Model based TI segmentation methods: Model-based TI segmentation techniques employees an acoustic modeling stage while trying to do phonetic segmentation. Modeling stage is done either according to supervised approach or unsupervised approach. Recently Self Supervised Learning algorithm was used for phoneme segmentation task[27].

In literatures, research studies proposed various types of modeling approaches, like generalized Gamma distribution model[18], graphical models[26] and Microcanonical Multiscale Formalism (MMF)[11] and Acoustic Segment Modeling (ASM)[6]. Supervised and unsupervised machine learning techniques like ANN, k-means[28] and Genetic Algorithm GA[24] were also used to learn the discriminative acoustic model.

Supervised segmentation techniques: Khanagha *et al*.[11] proposed the application of a totally novel approach, entitled the Microcanonical Multiscale Formalism (MMF) to speech analysis. MMF is based on estimating local scaling parameters that describe the inter-scale correlations at each point in the signal domain and provides efficient means for studying local non-linear dynamics of complex signals[11]. In this research, Khanagha introduced an efficient way for estimation these parameters and showed that they convey relevant information about local dynamics of the speech signal and thus can be used for the task of phonetic segmentation. Khanagha *et al*.[11] developed a two-stage segmentation algorithm: in the first step, he introduced a new dynamic programming technique to efficiently generate an initial list of phoneme-boundary candidates and in the second step, he used hypothesis testing to refine the initial list of candidates. Experiments on the full TIMIT database showed that the proposed algorithm was significantly more accurate than state-of the-art ones.

Inspired by the success of using Neural Networks in speech recognition, different studies[25, 29] considered applying them to phoneme segmentation task. Different types of ANN were investigated. Dinler *et al*.[25] and Wang *et al*.[6] suggested using Gated Recurrent Unit (GRU) recurrent neural networks, while Kreuk *et al*.[20] and Franke *et al*.[30] proposed using bidirectional LSTM (Long-Short Term Memory) network. Lu *et al*.[31] investigated the use of segmental Recurrent Neural Network (RNN) for feature extraction. Lee proposed using the cross-entropy loss with connectionist temporal classification loss in deep speech architecture for phoneme segmentation for the purpose of performing speech synthesis. Wang *et al*.[6] observed through experiments on the TIMIT corpus that GRU forget gate activations in trained recurrent acoustic neural networks correlate very well with phoneme which makes them preferable architecture for the task of boundary detection task. The advantage of both GRU and LSTM over standard recurrent neural networks RNN is their ability to incorporate long temporal context information and thus they give higher performance[25]. The GRU ensures the control of the information flow, similar to the LSTM unit but without a need to utilize a memory unit[25]. The GRU has a simpler structure compared to standard LSTM models and its popularity is gradually increasing[25].

Unsupervised segmentation techniques: Almpanidis and Kotropoulos[18] proposed a text-independent automatic phone segmentation algorithm based on the Bayesian Information Criterion for model parameter estimation. One of the main difficulties in phonetic segmentation is that it requires high resolution in the short-time analysis of the speech signal; while current analysis tools provides very limited information available in such a small scale. In order to tackle this issue, Almpanidis proposed modeling speech samples with the

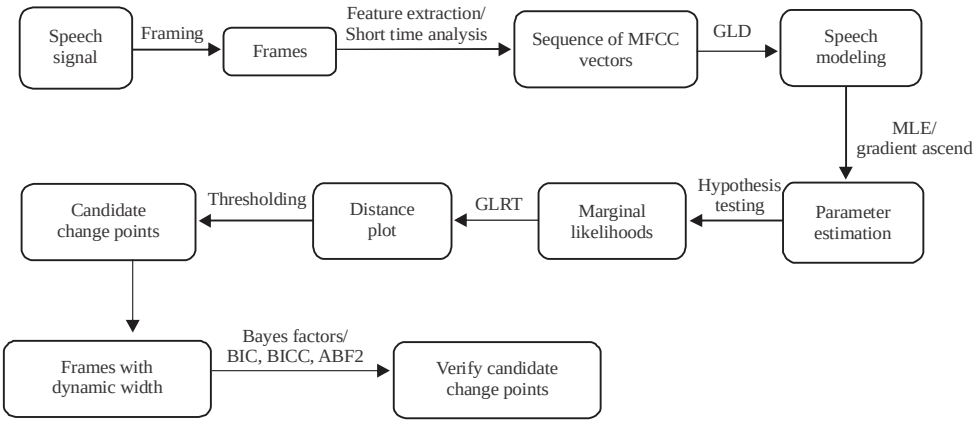

Fig. 4: Block diagram of the proposed TI segmentation system[18]

generalized Gamma distribution which is found to be more efficient than the Gaussian distribution[18]. He used a computationally inexpensive maximum likelihood approach for parameter estimation of the distribution and assessed the efficiency of the proposed algorithm on two different data sets. Results showed that the proposed technique yields significant performance improvement in noisy environments. The block diagram of the proposed system[18] is showed in Fig. 4.

In a later work, Almpanidis *et al*.[32] studied the efficiency of the phone boundary detection systems that employee entropy- and Bayesian-based model selection criteria in continuous speech based on the DISTBIC[33] hybrid segmentation algorithm. DISTBIC (DISTance and Bayesian Information Criterion (BIC)) is a hybrid technique that combines distance measures with BIC. It is a text-independent bottom-up approach that identifies sequential model changes by combining metric distances with statistical hypothesis testing[32]. Employing robust statistics and small sample corrections in the baseline DISTBIC algorithm, phone boundary detection accuracy is significantly improved while false alarms are reduced. Almpanidis noted that further improvement in segmentation accuracy was achieved by two additional steps: first considering how the model parameters are related in the probability density functions of the underlying hypothesis as well as in the model selection via the information complexity criterion and by second by employing M-estimators of the model parameters. Figure 4 depicts the block diagram of DISTBIC segmentation system[32].

Qiao *et al*.[34] formulated the segmentation problem into an optimization framework and developed an objective function which is the Summation of Squared Error (SSE) based on the Euclidean distance of cepstral features[34]. In a later work, Qiao and Minematsu[35] noted that it is unknown whether or not Euclidean distance verify the best metric to estimate the goodness of segmentations. Qiao *et al*.[34] studied the problem of learning a good metric to improve the performance of segmentation and proposed two criteria for learning metric: Minimum of Summation Variance (MSV) and Maximum of Discrimination Variance (MDV). The experimental results on TIMIT database showed that the use of learning metric increase segmentation performance. The best recall rate using the proposed learnt metric was 81.8% compared to 77.5%[35].

Qiao *et al*.[14] developed five different objective functions, namely Log Determinant (LD), Rate Distortion (RD), Bayesian Log Determinant (BLD), Mahalanobis Distance (MD) and Euclidean Distance (ED) objectives. He also introduced a time-constrained agglomerative clustering algorithm to find the optimal segmentations. Experiments on the TIMIT database showed that using the RD objective function achieves the best performance and that the proposed method outperforms the previous unsupervised segmentation methods (Fig. 5).

Acoustic Segment Modeling (ASM) is a common framework used for unsupervised acoustic modeling[6]. This approach consists of three stages, namely initial segmentation, segment labeling and iterative modeling[6]. In the initial segmentation stage continuous speech wave is divided into variable-length segments that have specific homogeneous acoustic properties. The speech segments are then clustered into a number of groups based on their acoustic similarities. Accordingly, each segment is mapped to one cluster label. Afterward the acoustic model for each cluster is estimated through an iterative process. Wang *et al*.[6] proposed a variation on this framework. He proposed using Gaussian Component Clustering (GCC) method and a Segment Clustering (SC) method for segment labeling. GCC applies spectral clustering on a set of Gaussian components while SC applies spectral clustering on speech segments. He examined the performances of the proposed ASM approaches in two applications: phonetic segments clustering on OGI

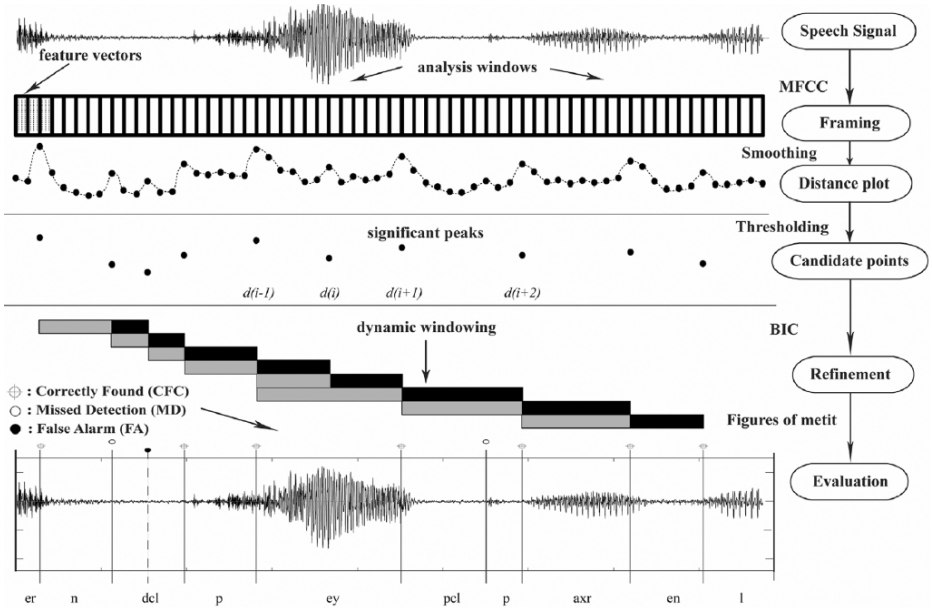

Fig. 5: Block diagram of DISTBIC phonemic segmentation algorithm[32]

multilingual telephone speech (OGI-MTS) corpus[26] and building a zero-resource Query-by-Example (QbyE) Spoken Term Detection (STD) application. Results showed that the proposed approach achieves good efficiency in both applications.

Chen *et al*.[13] proposed a new approach for text-independent Phoneme Segmentation at sampling point level. The algorithm consisted of two phases: Firstly, the voiced sections in speech data were detected using the information of vocal folds vibration contained in Electroglottograph (EGG). A Hilbert Envelope feature was adopted to achieve sampling point level detection accuracy. Secondly, the voiced sections and other sections were treated separately. Each voiced section was divided into several candidate phonemes using the Viterbi algorithm. Then adjacent candidate phonemes were merged based on a Hotellings T-square test method. For other sections, the unvoiced consonants were detected from silence based on a Singularity Exponent feature. Comparison experiments showed that the proposed method had better performance than the existing ones for a variety of tolerances and was more robust to noise.

Kamper *et al*.[28] introduced a new approach based on k-means clustering algorithm as an unsupervised learning algorithm for word and phoneme boundaries detection. The new approach entitled Embedded Segmental k-Means model (ES-KMeans) gives similar scores to the Bayesian model while being 5 times faster with fewer hyper-parameters[28]. Experiments also show that ES-KMeans scales to larger corpora by applying it to the 5 languages of the Zero Resource Speech Challenge 2017[28].

Kreuk *et al*.[15] proposed a Self-Supervised representation Learning (SSL) model for phoneme boundary detection. They proposed learning a feature representation from the raw waveform to identify spectral changes that match phoneme boundaries accurately. For this task, they designed a Convolutional Neural Network (CNN) to distinguish between pairs of adjacent frames and pairs of random distractor pairs. At test time, a peak detection algorithm is applied over the model outputs to produce the final boundaries[15]. Results show that the proposed SSL technique surpasses other unsupervised segmentation techniques.

Model-free (Blind) TI segmentation methods: Model-free phonetic segmentation methods (also called metric-based or blind methods) does not incorporate any modeling strategy in tackling the segmentation task, instead they rely on distance measures of the spectral changes among consecutive speech frames. These method uses the signal characteristics extracted in a signal analysis stage and a collection of thresholds to segment the signal[36]. The main issue with this approach is the difficulty to determine the optimal threshold.

Dusan and Rabiner[23] investigated the use of spectral transition in segmentation, as he found high correlation between the maximum of the spectral transition and phoneme boundaries. The proposed method detects phoneme boundaries by looking for peaks in a spectral transition metric. Results showed an accuracy of 84.6% at a 20 msec threshold TIMIT dataset while no other performance metric was reported.

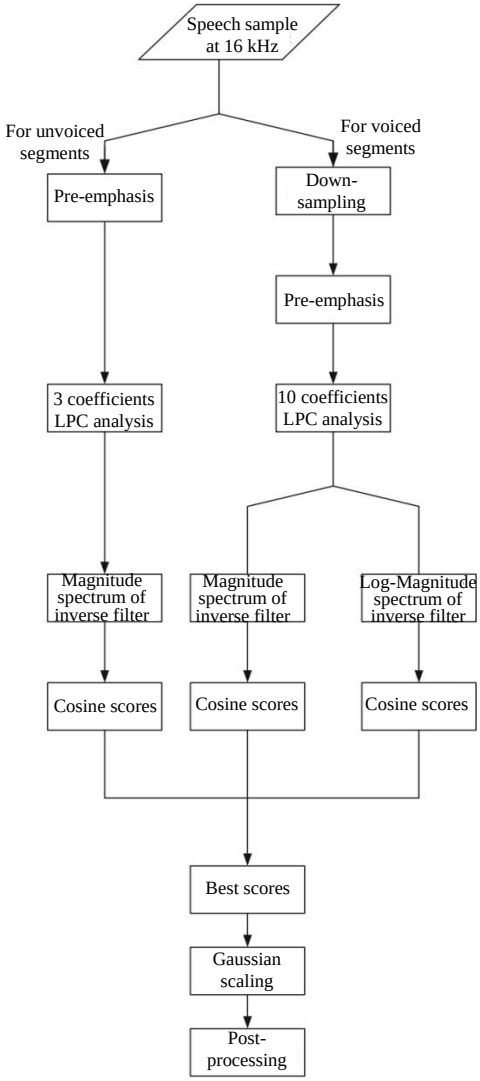

Fig. 6: Flowchart of proposed scheme for automatic segmentation of speech based on FICV[24]

Ziolko *et al*.[37, 38] proposed a new phoneme segmentation method based on the analysis of discrete wavelet transform spectra. His method relies on the values of power envelopes and their first derivatives for six frequency sub-bands. Specific scenarios that are typical for phoneme boundaries were searched for.

Discrete times with such events are noted and graded using a distribution-like event function which represent the change of the energy distribution in the frequency domain. The final decision on localization of boundaries was taken by analysis of the event function. Boundaries are, therefore, extracted using information from all sub-bands. The method was developed on a small set of Polish hand segmented words and tested on another large corpus containing 16425 utterances. A recall and precision measure specifically designed to measure the quality of speech segmentation was adapted by using fuzzy sets. Ziolko *et al*.[38] showed that results with F1-score equal to 72.49% were obtained.

Javed *et al*.[39] proposed a strategy driven by cosine distance similarity scores for identifying phoneme boundaries. The proposed strategy helped in the selection of appropriate feature extraction technique for speech segmentation applications. After assessing various state-of the-art speech processing techniques, a new combination of Forward and Inverse Characteristics of Vocal tract (FICV) was introduced. Experimental results on Classical Arabic dataset showed that proposed technique has total error rate of 14.48% while the accuracy is 85.2% within 10 msec alignment error. When compared with the existing state-of-the-art technique, the proposed technique outperforms by 12.29 and 22.73% in terms of error rates and alignment accuracies, respectively[39]. The block diagram of the proposed system is depicted in Fig. 6[39].

Ramteke and Koolagudi[19] noted that in a well spoken word, phonemes can be characterized by the changes observed in speech waveform. To get phoneme boundaries, Ramteke studied the signal level properties of speech waveform, i.e., changes in the waveform during transformation from one phoneme to the other. He addressed the problem of phoneme level segmentation from two aspects: segmentation of phonemes between voiced and unvoiced portions and segmentation of phonemes within voiced and unvoiced regions. He used pitch and zero-frequency filter signal to get the region of change from voiced to unvoiced and vice versa.

The segmentation of phoneme boundaries within voiced and unvoiced regions are approximated using the properties of power spectrum of correlation of adjacent frames of the signal. Finally, he proposed a finite set of rules on the variations observed in the power spectrum during phoneme transitions.

The segmentation results of both approaches are combined to get the final phoneme boundaries. Three databases were used to test the proposed approach; an accuracy of 95.40, 96.87 and 96.12% is achieved within the tolerance range of 10 msec respectively. The proposed system block diagram is depicted in Fig. 7[19].

PERFORMANCE COMPARISON

In Table 1, we provide a detailed performance comparison of most of the TI segmentation methods in terms of hit rate.

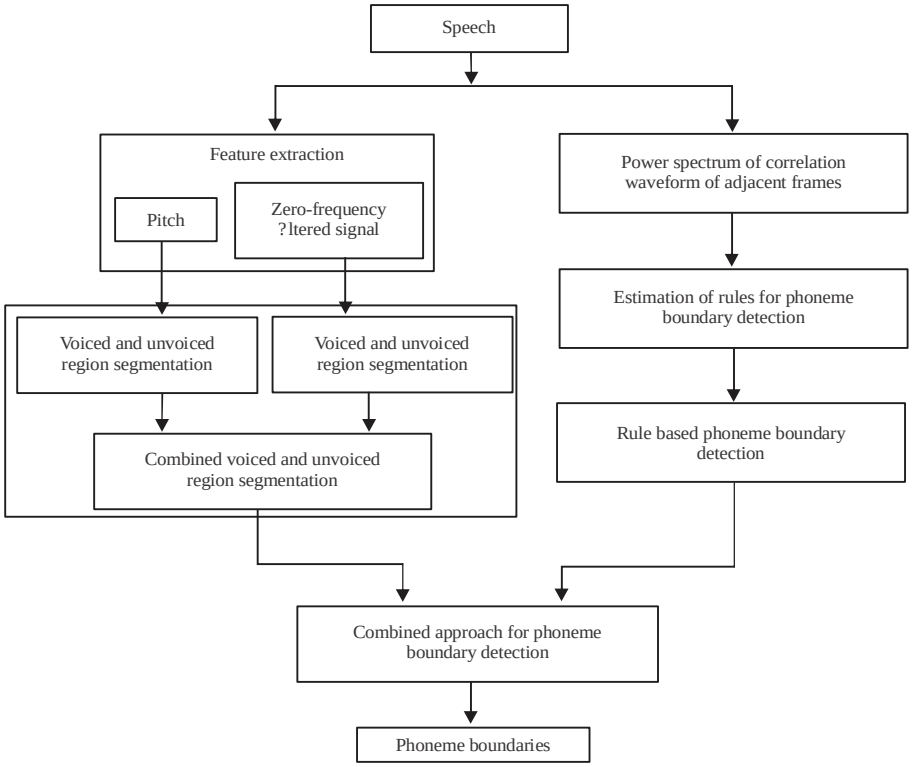

Fig. 7: Phoneme boundary detection: a rule based approach[19]

Table 1: Performance comparison of most of the TI segmentation methods

References	B/S/U/SS[5]	Dataset	Features	Classifier	Hit rate (Recall)
Dusan and Rabiner[23]	B	TIMIT database	MFCCs	-	Tolerance≤10 70.00%
Zolko *et al*.[38]	U	Polish speech recordings Corpora'97 database	DWT	-	Phoneme recognition rate of 81.00% at 25 msec
Khanagha *et al*.[11]	S	TIMIT database	Microcanonical Multiscale Formalism (MMF)	Piece-wise-linear approximation followed by Log-Likelihood Ratio Test (LLRT)	Tolerance≤10 msec: hit rate = 53.16%
Chen *et al*.[13]	U	Database of German Emotional Speech (Berlin EmoDB)	MFCCs	-	Tolerance≤20 msec: Hit rate = 82.00% False alarm = 20% F1 = 81.00%
Wang *et al*.[6]	S	TIMIT database	MFCCs, ΔMFCCs, ΔΔMFCCs,	Autoencoder Gated Recurrent Neural Network AE-GRNN	R-value: 83.61%
Ramteke and Koolagudi[19]	B	TIMIT Coupus; IIIT-H Indic speech databases-Marathi; IIIT-H Indic speech databases-Hindi	Voiced and unvoiced segmentation: Energy of Zero Frequency Filter signal, Pitch; Phoneme segmentation within voiced and unvoiced regions: Rules based on the nature of Power Spectrum of Correlation waveform of consecutive frames	-	Tolerance≤10 msec: TIMIT Corpus: 97.00%; IIIT-H Indic speech databases-Marathi: 98.00%; IIIT-H Indic speech databases Hindi: 98.00%
Kreuk *et al*.[20]	S	TIMIT Buckeye	Segmental feature	Bidirectional Recurrent Network BI-RNN	Tolerance≤20 msec 90.46%
Kreuk *et al*.[15]	SSL	TIMIT database Buckeye	Learned featured from raw waveform	CNN	Tolerance≤20 msec 83.55%

[5]B = Blind; U = Unsupervised; S = Supervised; SS = Self Supervised

CONCLUSION

In this study, we exposed an in-depth survey of the different TI phonetic segmentation algorithms that exist in literature so far. These techniques are classified into blind text independent techniques, supervised text independent techniques and self-supervised learning segmentation techniques. We provided a detailed performance comparison showing that the latest state-of rate TI phonetic segmentation that employs a rule based approach proposed by Ramteke and Koolagudi[19], though being a blind segmentation technique, still it achieves the best performance in terms of accuracy of all the other explicit techniques on TIMIT corpus.

We have noted that with the exception of the research done by Zolko *et al*.[24] who used wavelet for feature extraction, almost all the reviewed studies used MFCCs or Pitch and Zero crossing acoustic features, though wavelet based features has achieved better performance in phoneme recognition task[27].

REFERENCES

01. Hemert, J.P.V., 1991. Automatic segmentation of speech. IEEE. Trans. Signal Process., 39: 1008-1012.
02. Brugnara, F., D. Falavigna and M. Omologo, 1993. Automatic segmentation and labeling of speech based on hidden Markov models. Speech Commun., 12: 357-370.
03. Glass, J.R., 2003. A probabilistic framework for segment-based speech recognition. Comput. Speech Lang., 17: 137-152.
04. Chappell, D.T. and J.H. Hansen, 2002. A comparison of spectral smoothing methods for segment concatenation based speech synthesis. Speech Commun., 36: 343-373.
05. Adell, J. and A. Bonafonte, 2004. Towards Phone Segmentation for Concatenative Speech Synthesis. Proceedings of the 5th ISCA Workshop on Speech Synthesis, June 14-16, 2004, Institute of Singapore Chartered Accountants, Pittsburgh, Pennsylvania, pp: 139-144.
06. Wang, H., T. Lee, C.C. Leung, B. Ma and H. Li, 2015. Acoustic segment modeling with spectral clustering methods. IEEE/ACM. Trans. Audio Speech Lang. Process., 23`: 264-277.
07. Hosom, J.P., 2009. Speaker-independent phoneme alignment using transition-dependent states. Speech Commun., 51: 352-368.
08. Ljolje, A., J. Hirschberg and J.P.H.V. Santen, 1997. Automatic Speech Segmentation for Concatenative Inventory Selection. In: Progress in Speech Synthesis, Santen, J.P.H.V., J.P. Olive, R.W. Sproat and J. Hirschberg (Eds.)., Springer, Berlin, Germany, pp: 305-311.
09. Pellom, B.L. and J.H. Hansen, 1998. Automatic segmentation of speech recorded in unknown noisy channel characteristics. Speech Commun., 25: 97-116.
10. Esposito, A. and G. Aversano, 2004. Text Independent Methods for Speech Segmentation. In: Nonlinear Speech Modeling and Applications, Chollet, G., A. Esposito, M. Faundez-Zanuy and M. Marinaro (Eds.)., Springer, Berlin, Germany, pp: 261-290.
11. Khanagha, V., K. Daoudi, O. Pont and H. Yahia, 2014. Phonetic segmentation of speech signal using local singularity analysis. Digital Signal Process., 35: 86-94.
12. Toledano, D.T., L.A.H. Gomez and L.V. Grande, 2003. Automatic phonetic segmentation. IEEE. Trans. Speech Audio Process., 11: 617-625.
13. Chen, L., X. Mao and H. Yan, 2016. Text-independent phoneme segmentation combining egg and speech data. IEEE/ACM. Trans. Audio Speech Langu. Process., 24: 1029-1037.
14. Qiao, Y., D. Luo and N. Minematsu, 2013. Unsupervised optimal phoneme segmentation: Theory and experimental evaluation. IET Signal Process., 7: 577-586.
15. Kreuk, F., J. Keshet and Y. Adi, 2020. Self-supervised contrastive learning for unsupervised phoneme segmentation. Proc. Interspeech, 1: 3700-3704.
16. Mporas, I., T. Ganchev and N. Fakotakis, 2010. Speech segmentation using regression fusion of boundary predictions. Comput. Speech Lang., 24: 273-288.
17. Almpanidis, G. and C. Kotropoulos, 2008. Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion. Speech Commun., 50: 38-55.
18. Rasanen, O.J., U.K. Laine and T. Altosaar, 2009. An improved speech segmentation quality measure: The R-value. Proceedings of the 10th Annual Conference of the International Speech Communication Association, September 6-10, 2009, Interspeech, Shanghai, China, pp: 1851-1854.
19. Ramteke, P.B. and S.G. Koolagudi, 2019. Phoneme boundary detection from speech: A rule based approach. Speech Commun., 107: 1-17.
20. Kreuk, F., Y. Sheena, J. Keshet and Y. Adi, 2020. Phoneme boundary detection using learnable segmental features. Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-8, 2020, IEEE, Barcelona, Spain, pp: 8089-8093.

21. Sarma, B.D. and S.M. Prasanna, 2018. Acoustic-phonetic analysis for speech recognition: A review. IETE. Tech. Rev., 35: 305-327.
22. Hoang, D.T. and H.C. Wang, 2015. Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. J. Acoust. Soc. Am., 137: 797-805.
23. Dusan, S. and L. Rabiner, 2006. On the relation between maximum spectral transition positions and phone boundaries. Proceedings of the 9th International Conference on Spoken Language Processing, September 17-21, 2006, Interspeech, Rutgers University, New Jersey, USA., pp: 645-648.
24. Ziolko, M., J. Galka, B. Ziolko and T. Drwiega, 2010. Perceptual wavelet decomposition for speech segmentation. Proceedings of the 11th Annual Conference of the International Speech Communication Association, September 26-30, 2010, ISCA, Makuhari, Chiba, Japan, pp: 2234-2237.
25. Dinler, O.B. and N. Aydin, 2020. An optimal feature parameter set based on gated recurrent unit recurrent neural networks for speech segment detection. Appl. Sci., Vol. 10, 10.3390/app10041273
26. Teng, P., X. Liu and Y. Jia, 2013. Text-Independent Phoneme Segmentation Via Learning Critical Acoustic Change Points. In: Intelligent Science and Big Data Engineering, Sun, C., F. Fang, Z. Zhou, W. Yang and Z. Liu (Eds.)., Springer, Berlin, Germany, pp: 54-61.
27. Sahu, P.K., A. Biswas, A. Bhowmick and M. Chandra, 2014. Auditory ERB like admissible wavelet packet features for TIMIT phoneme recognition. Eng. Sci. Technol. Int. J., 17: 145-151.
28. Kamper, H., K. Livescu and S. Goldwater, 2017. An embedded segmental k-means model for unsupervised segmentation and clustering of speech. Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 16-20, 2017, IEEE, Okinawa, Japan, pp: 719-726.
29. Chih-Kuan, Y., J. Chen, C. Yu and D. Yu, 2019. Unsupervised speech recognition via segmental empirical output distribution matching. Proceedings of the International Conference on Representation Learning, (ICLR), May 6-9, 2019, New Orleans, Louisiana, pp: 1-14.
30. Franke, J., M. Mueller, F. Hamlaoui, S. Stueker and A. Waibel, 2016. Phoneme boundary detection using deep bidirectional LSTMS. Proceedings of the Speech Communication: 12. ITG Symposium, October 5-7, 2016, VDE, Paderborn, Germany, pp: 1-5.
31. Lu, L., L. Kong, C. Dyer, N.A. Smith and S. Renals, 2016. Segmental recurrent neural networks for end-to-end speech recognition. Comput. Lang., Vol. 1,
32. Almpanidis, G., M. Kotti and C. Kotropoulos, 2019. Robust detection of phone boundaries using model selection criteria with few observations. IEEE. Trans. Audio Speech Lang. Process., 17: 287-298.
33. Delacourt, P. and C.J. Wellekens, 2000. DISTBIC: A speaker-based segmentation for audio data indexing. Speech Commun., 32: 111-126.
34. Qiao, Y., N. Shimomura and N. Minematsu, 2008. Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, March 31-4, 2008, IEEE, Las Vegas, New York, USA., pp: 3989-3992.
35. Qiao, Y. and N. Minematsu, 2008. Metric learning for unsupervised phoneme segmentation. Proceedings of the 9th Annual Conference of the International Speech Communication Association, September 22-26, 2008, Interspeech, Tokyo, Japan, pp: 1060-1063.
36. Frihia, H. and H. Bahi, 2017. HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications. Int. J. Speech Technol., 20: 563-573.
37. Ziolko, B., S. Manandhar, R.C. Wilson and M. Ziolko, 2006. Wavelet method of speech segmentation. Proceedings of the 2006 14th European Signal Processing Conference, September 4-8, 2006, IEEE, Florence, Italy, pp: 1-5.
38. Ziolko, B., S. Manandhar, R.C. Wilson and M. Ziolko, 2011. Phoneme segmentation based on wavelet spectra analysis. Arc. Acoust., 36: 29-47.
39. Javed, M., M.M.A. Baig and S.A. Qazi, 2020. Unsupervised phonetic segmentation of classical Arabic speech using forward and inverse characteristics of the vocal tract. Arab. J. Sci. Eng., 45: 1581-1597.