Research Journal of
Applied Science
MEDWELL PUBLICATIONS

An Inclusive Survey for Text Dependent 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-dependent, fusion, predictive models, speech parameterization, HMM

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

Page No.: 65-74
Volume: 16, Issue 2, 2021
ISSN: 1815-932x
Research Journal of Applied Sciences

Abstract: Speech segmentation is the process of breaking speech signal into distinct acoustic blocks that could be words, syllabus or phonemes. Phonetic segmentation is about finding the exact boundaries for the different phonemes that composes a specific speech signal. Phonetic segmentation is crucial for many applications basically speech recognition ASR and speech to text systems STT as ASR needs phonetically transcribed training corpus, STT needs phoneme database. 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. In this study, we present a thorough survey of the different algorithm and techniques proposed so far for solving the problem of text-dependent 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 algorithms development of 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 and 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 three broad categories: unsupervised techniques, Self-Supervised Learning (SSL) techniques and supervised segmentation techniques[15]. In supervised methods (also called model-based) a training stage is needed to learn an acoustic model that can help in discriminating borders from non-borders segments. After learning a model, the segmentation process is done through binary classification. These techniques need manually segmented dataset to train the model[16, 17]. The unsupervised methods (also called blind, or model-free) overcome this problem 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, 17]. 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]. Figure 1 depicts the classification of different phonetic segmentation systems.

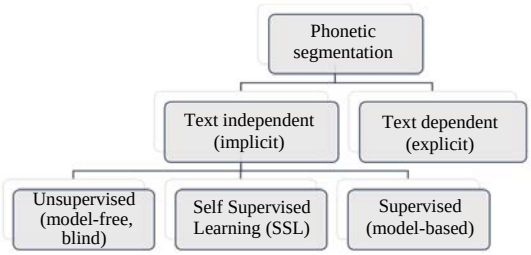

Fig. 1: Classification of phonetic segmentation techniques

In this research, we present an inclusive survey of the text-dependent phonetic segmentation techniques.

TD SEGMENTATION ALGORITHMS PERFORMANCE METRICS

Different metrics are used to evaluate the segmentation algorithms performance. In the text dependent phonetic segmentation techniques, there are three metrics Accuracy, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The number of annotated segments is exactly equal to that of the manually segmented corresponding signal, thus, we only compare the boundaries of the manual segmentation to the boundaries discovered by the algorithm.

Boundaries that are deviated more than a specific threshold (Most commonly the threshold is equal to 20 msec and accuracy for different thresholds is reported in literature) are reported as errors and accordingly accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{Correct boudaries}}{\text{Total actual boudaries}} \times 100\% \quad (1)$$

The Mean Absolute Error (MAE) is the average of absolute deviation (in millisecond) between the discovered boundaries and the manually marked ones. This deviation measure can be reported in terms of Root Mean Squared Error (RMSE).

Figure 2 summaries the metrics used in the assessment of different reviewed TD segmentation algorithms.

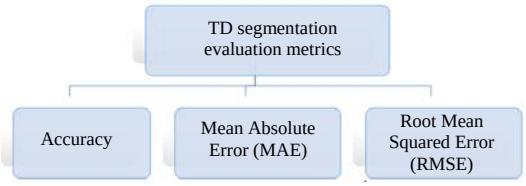

Fig. 2: Different metrics used for evaluating TD phonetic segmentation techniques

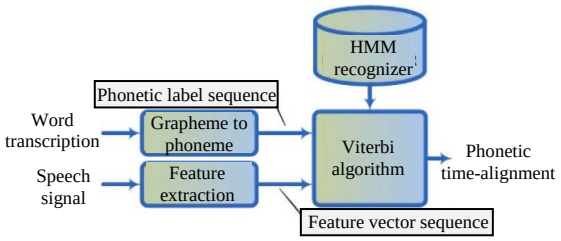

Fig. 3: HMM-based phonetic segmentation[18]

TEXT DEPENDENT (SUPERVISED) PHONETIC SEGMENTATION TECHNIQUES

The most frequently used approach for text dependent segmentation is based on HMM phone modeling[2, 12, 18]. This approach is inspired from the well-known structure of ASR systems. In this method each speech waveform is initially spliced into a sequence of overlapping frames and then feature extraction is applied to each frame to produce feature vectors using a specific speech parameterization technique. Afterwards, a set of HMM phone models manually trained using annotated speech waveforms is utilized to detect the corresponding phoneme of each feature vectors. Finally, phonetic boundaries are detected as a byproduct of the phoneme recognition task. Each phone label sequence is force-aligned against the corresponding feature vector sequence and the phone model set, through the Viterbi algorithm[8]. The block diagram of the HMM-based phonetic segmentation system for the text dependent (explicit) case is shown in Fig. 3[18].

Template matching using Dynamic Time Warping (DTW) was also proposed for forced-alignment but it gave inferior performance.

Reported results of the traditional phonetic segmentation techniques (HMM-based, DWT) in TTS application are still far from being satisfactory[19]. Segmentation accuracy needed for TTS applications are higher than for ASR applications. Since, ASR systems aim to find the correct annotation sequence of the speech waveform and this does not demand an accurate placement of phone boundaries as the case for TTS systems[10]. That is why various studies are proposed to enhance the phonetic segmentation accuracy of the existing segmentation systems.

Recent studies for enhancing the segmentation accuracy in text dependent techniques can be classified into two groups based on: modifying the structure of acoustic models and post-processing techniques which try to refine the preliminary boundaries produced by an existing segmentation system[20] as shown in Fig. 4.

Post-processing techniques: Post-processing techniques are based on refining the initial segmentation results through different procedures. The main idea behind these methods is inspired from the manual segmentation process of human labelers. They first listen to the speech signal to get a rough boundary (baseline segmentation system), then examine the spectrogram or waveform in more detail to identify the accurate boundary (post-processing: boundary refinement). In literature, different post-processing techniques have been proposed. They can be categorized into four different groups: statistical correction, fusion, predictive models and hybrid.

Statistical correction: The main idea in this technique is to calculate a statistical average of the error produced by the segmentation system and trying to compensate this error. Figure 5 depicts the block diagram of such boundary statistical correction system[21].

Toledano *et al.*[12] proposed a new statistical correction technique entitled Statistical Correction of Context Dependent Boundary Marks (SCCDBM) for handling the systematic errors produced by context-dependent HMMs. This new technique (SCCDBM) includes two steps: training phase and boundary phase. In the training phase, the statistical averages of the boundaries error are estimated. In the boundary correction phase, the phone boundaries are moved according to those estimated averages. Experiments on Castilian Spanish corpora showed that the proposed SCCDBM system achieves an accuracy of 91.18% in tolerance region of 20 msec indicating a notable increase compared with the baseline context independent and context dependent HMMs segmentation whose accuracy were 82.70 and 79.41%, respectively.

Toledano *et al.*[12] proposed a new statistical correction technique entitled Statistical Correction of Context Dependent Boundary Marks (SCCDBM) for handling the systematic errors produced by context-dependent HMMs. This new technique (SCCDBM) includes two steps: training phase and boundary phase. In the training phase, the statistical averages of the boundaries error are estimated. In the boundary correction phase, the phone boundaries are moved according to those estimated averages. Experiments on Castilian Spanish corpora showed that the proposed SCCDBM system achieves an accuracy of 91.18% in tolerance region of 20 msec indicating a notable increase compared with the baseline

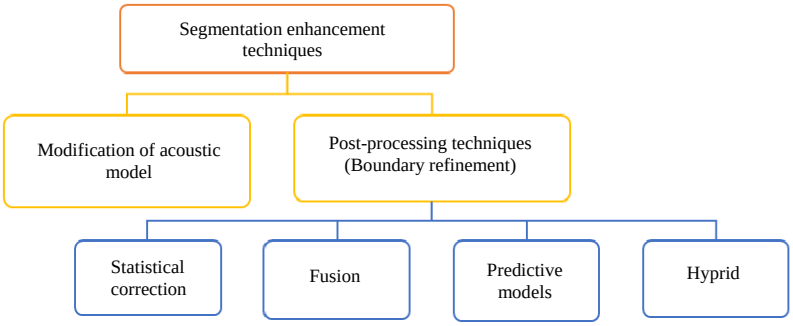

Fig. 4: Different techniques for enhancing text dependent segmentation

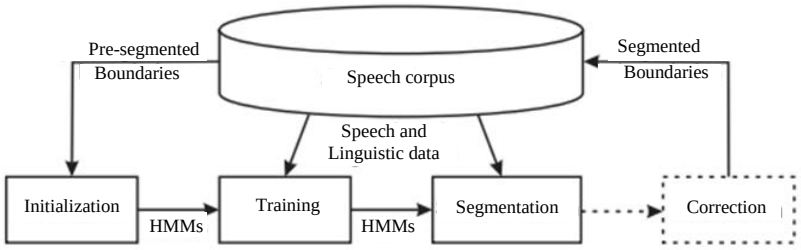

Fig. 5: Simplified illustration of the statistical approach to speech segmentation with optional correction[21]

context independent and context dependent HMMs segmentation whose accuracy were 82.70 and 79.41%, respectively.

Matousek *et al*.[21] proposed a Boundary-Specific Statistical correction (BSC) technique for enhancing the segmentation results of the HMM based segmentation system. The proposed technique comprises two passes. In the first pass all boundaries in the training dataset are corrected by shifting the boundary with respect to the boundary-specific average deviation. In the second pass, the corrected segmented dataset is used as the input for the segmentation of test dataset. Experiments on Czech corpus composed of 5000 hand-labeled sentences showed that the proposed refinement method achieves an accuracy of 96% in tolerance region of 20 msec.

Fusion: Fusion techniques are based on combining the segmentation results of different Automatic Segmentation Machines (ASMs) to produce final boundaries time-marks. This is done through either selecting the most appropriate boundary from segmentation results of the different ASMs depending on the phonetic context[22] or through a weighted summation of segmentation results[19].

Park and Kim[22] proposed a new approach to improve the performance of automatic speech segmentation techniques for concatenative TTS synthesis. Given multiple ASMs, multiple boundary sets denoting the collection of boundary time marks are produced by multiple ASMs which adopt different methods from each other. Then, for each boundary type, the candidate selector chooses the best time mark among the boundaries provided by the multiple ASMs. For each boundary type observed in the training database, the average time differences between the target time marks and those provided by the ASMs are computed. Then the ASM with the minimal error is selected as the winner for the given boundary type[22]. The proposed system is entitled Automatic Segmentation by Boundary Type Candidate Selection (ASBTCS).

Jarifi *et al*.[23] analyzed three automatic segmentation algorithms that he proposed to combine into one fusion system. The first algorithm is segmentation by HMM. The second one is entitled refinement by boundary model, where a Gaussian Mixture Model (GMM) of each boundary is used to improve the initial segmentation performed by HMM. The third algorithm is a slightly modified version of Brandt's Generalized Likelihood Ratio (GLR) method; its goal is to detect signal discontinuities in the vicinity of the HMM boundaries. Experimental results on a specific dataset showed that the most accurate fusion method, called optimal fusion by soft supervision, reduces by 25.5, 60 and 75%, the number of segmentation errors made by refinement by boundary model, standard HMM segmentation and Brandt's GLR method, respectively. As for TTS applications, subjective listening tests showed that the quality of the synthetic speech obtained when the speech corpus is segmented by optimal fusion by soft supervision is close to that obtained when the same corpus is manually segmented.

Predictive models: The main idea in this technique is to learn a predictive model from training dataset for phoneme boundaries, that can be applied later to predict the most likely one on testing dataset. Different predictive models from machine learning were employed in literature, like linear and non-linear regression[24], k-Nearest Neighborhood (kNN)[25], Multiple-Layer Perception (MLP)[26] and support vector machine SVM[27].

Park and Kim[19] proposed using a spectral transition measure to find the maximum spectral distortions as a predicative model for boundary refinement.

Lee[26] considered the use of MLP for refinement process. He proposed training a MLP for each kind of phone transition groups and used the obtained model for boundary correction. The optimum partitioning of the entire phonetic transition space and the corresponding MLPs were constructed with the objective of minimizing the overall deviation from the manually marked boundaries. The experimental results showed that >93% of all phone boundaries have a boundary deviation from a reference position <20 msec. Lee[26] also confirmed that the synthesis speech produced using the database constructed by the proposed method was perceptually comparable to a that produced using hand-labeled database, based on subjective listening tests.

The methods presented in[28, 19, 27, 29] are all based on using SVM for boundary refinement. The feature vectors combining both spectral and prosodic information are used to represent the frames around the preliminary boundary from forced alignment and then SVMs are used to identify the most probable boundary.

Frihia and Bahi[16] proposed a combination of Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) to segment and label the speech waveform into phoneme units. HMMs generate the sequence of phonemes and their boudaries; the multi-class SVM refines the boundaries and corrects the labels. The obtained segmented and labelled units may serve as a training set for speech recognition applications. The HMM/SVM segmentation algorithm was assessed using both the hit rate and the Word Error Rate (WER); the resulting scores were compared to those provided by the manual segmentation and to those provided by the well-known embedded learning algorithm. Experiments on Arab Phone corpus showed that the speech recognizer built upon the HMM/SVM segmentation outperforms in terms of WER the one built upon the Embedded Learning (EL) segmentation of about 0.05%, even in noisy background. In terms of accuracy, results showed that using SVM increases the quality of segmentation compared to the segmentation using EL within tolerances 20, 35 and 45 msec but not within tolerance of 10 msec.

Hybrid: Hybrid post processing techniques refer to methods that combine two or more post-processing techniques. Park and Kim[19] proposed an approach that combines both fusion and statistical correction techniques. He proposed using multiple independent biased corrected ASMs to produce a nal boundary time-mark. A training procedure using manually segmented dataset is utilized to obtain the bias and weight parameters. The bias and weight parameters are calculated by averaging the errors of each phonetic context in case the cost function is a squared error. Afterward the bias parameters are xed and the weight parameters are calculated through a gradient projection optimization method with a set of constraints imposed on the weight parameter space. A decision tree which clusters all the phonetic contexts was utilized to deal with the unseen phonetic contexts. This proposed system is entitled Automatic Segmentation by Weighted Sum of Multiple Bias-Corrected Results (ASWSBC). Experimental results indicated that ASWSBC achieves a segmentation accuracy of 97.07% for a 20 msec threshold[19].

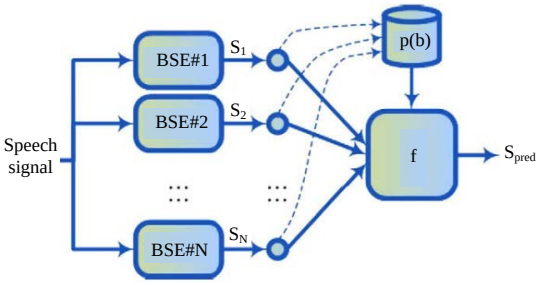

Fig. 6: Block diagram of the regression fusion of BSEs[30]

Lin and Jang[29] and Stolcke *et al*.[31] proposed a different hybrid scheme that combines both predictive models and fusion techniques together.

Mporas *et al*.[30] studied a number of linear and non-linear regression methods used for combining multiple phonetic boundary predictions obtained from different Baseline Segmentations Engines (BSE). The proposed fusion schemes were independent of the implementation of the individual segmentation engines as well as from their number. Mporas used 112 speech segmentation engines based on HMMs. He relied on sixteen different HMMs setups and on seven speech parameterization techniques. Experimental results on the phonetic segmentation task of TIMIT database showed that the support vector regression scheme achieves more accurate predictions compared to other fusion schemes. Figure 6 depicts the block diagram of the proposed system.

Lin and Jang[29] introduced the concept of a Score Predictive Model (SPM) that can re ne the phoneme boundaries of a fusion system obtained by HMM and DTW for a Mandarin singing voice corpus. Several experiments with different settings, including the use of different initial estimates, different acoustic features and various regression approaches were designed to verify the feasibility of the proposed approach. Experimental results demonstrate that the proposed SPM is able to effectively refine the results of the HMM and DTW[19].

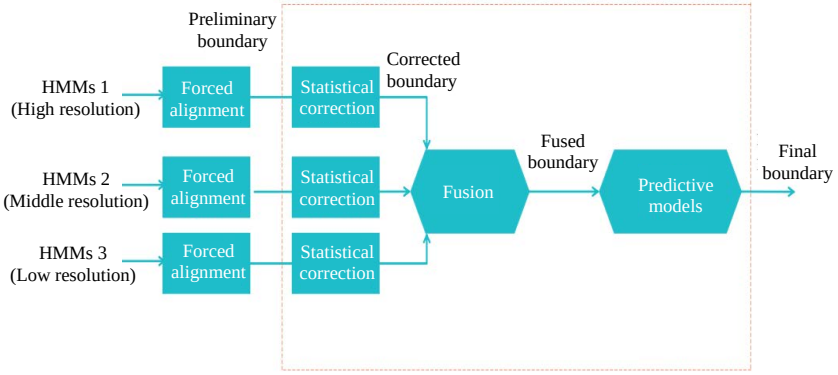

Fig. 7: Hybrid boundary refinement scheme[20]

Stolcke *et al*.[31] proposed a fusion system that combines the boundary estimates of multiple acoustic front-ends that uses different speech parameterization techniques. He studied two predictive models as boundary correction models: Neural Networks (NN) and regression trees. The proposed fusion is done by averaging the results of the boundary corrected estimates of each acoustic front-end using neural network which found to achieve the highest improvement among 3 other different fusion scenarios[31].

Zhao *et al*.[20] proposed a hybrid refinement scheme based on combining all the three different post-processing techniques previously exposed. A statistical method based on state-level correction was proposed to improve the segmentation results. A multi-resolution fusion process was proposed to study the stepsize effects and combined segmentations given by different HMMsto improve the accuracy. Then, predictive models with a smaller step size were designed to further refine the phone boundaries. By applying the hybrid refinement scheme on a well-known corpus, significant improvements of segmentation results were observed, in terms of segmentation accuracy with different tolerances MAE and RMSE. The proposed system block diagram is depicted in Fig. 7.

Furthermore, a scenario of cross-corpora segmentation was examined. Automatic phonetic segmentation is performed on TIMIT corpus using the segmentation system trained on a standard corpus WSJCAM0 (Cambridge wall street journal), a British version of WSJ by applying the proposed refinement scheme. Experimental results showed that the proposed refinement procedure can generate segmentation results comparable to those given by well-trained acoustic models obtained from the new corpus.

Modification of acoustic models: Another trend to improve phonetic segmentation accuracy is based on modifying the acoustic model. In literature, different modifications have been proposed. Mporas *et al*.[30] proposed an efficient Viterbi-based segmentation scheme, using multiple speech Fourier-based and wavelet-based speech parameterization techniques, in the acoustic modeling stage with respect to the phoneme boundary type as shown in Fig. 8.

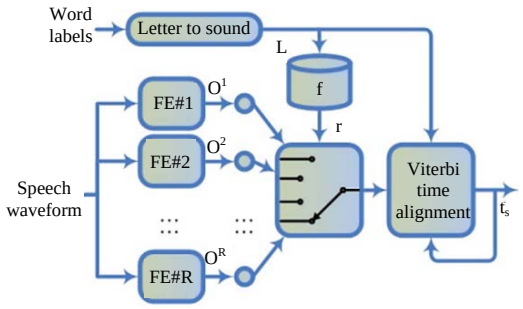

Fig. 8: Phonetic segmentation system using multiple speech features VSMSF[30]

The proposed system entitled Viterbi-based phoneme Segmentation with Multiple Speech Features (VSMSF) utilizes for each observation, the most accurate phonetic boundary prediction (f) which is obtained through the most appropriate among all speech features extraction (FE) as the initial point for the prediction of the next boundary position. The proposed method was evaluated on the TIMIT database, employing several speech parameterization techniques, like Fourier-based and wavelet-based.

Adell *et al*.[32] proposed applying Dynamic Time Warping (DTW) in combination with an acoustic clustering method to produce more accurate phonetic boundaries for TTS systems. Akdemir and Cilogu[28] proposed a new HMM topology that includes a special state for modeling the boundary, with one frame duration only. This topology is left to right and has three states. These states associated with the left phoneme class at the boundary, the single boundary frame and the right phoneme class. Phonemes are grouped into 10 classes with the addition of special classes for “breath” and “silence”, the boundary models are developed for each

class-to-class phoneme transition, resulting in boundary models. Keshet *et al*.[33] proposed using supervised learning algorithm to learn an alignment function which can be casted as a relaxed Support Vector Machine (SVM) optimization problem. Keshet proposed an iterative algorithm to solve. Results showed an improvement in segmentation accuracy compared with the standard HMM based method[33].

Hosom[7] proposed a modified HMM system employing Artificial Neural Network (ANN) to calculate probabilities. He proposed adding four energy-based features to the standard acoustic cepstral feature. Using probabilities of a state transition given an observation and a computation of (context-dependent) phoneme-level probability instead of phoneme probability into the HMM/AAN hybrid system. The (context-dependent) phoneme-level is a combination of probabilities of three distinctive phonetic features (manner of articulation, tongue position and height of tongue body) instead of phoneme-level probabilities.

Kim *et al*.[34] proposed a Large Margin Discriminative Semi-Markov Model (LMSMM) for phonetic recognition. The Hidden Markov Model (HMM) framework that is often used for phonetic recognition assumes only local statistical dependencies between adjacent observations. HMM is also used to predict a label for each observation without explicit phone segmentation. However, the Semi-Markov Model (SMM) framework allows simultaneous segmentation and labeling of sequential data based on a segment-based Markovian structure that assumes statistical dependencies among all the observations within a phone segment. For phonetic recognition which is inherently a joint segmentation and labeling problem, the SMM framework has the potential to perform better than the HMM framework at the expense of slight increase in computational complexity[34]. The SMM framework considered by Kim *et al*.[34] is based on a non-probabilistic discriminant function that is linear in the joint feature map which attempts to capture long-range statistical dependencies among observations. The parameters of the discriminant function were estimated by a large margin learning framework for structured prediction. The parameter estimation problem in hand lead to an optimization problem with many margin constraints and this constrained optimization problem was solved using a stochastic gradient descent algorithm[34]. Experimental results showed that the proposed LMSMM outperformed the large margin discriminative HMM in the TIMIT phonetic recognition task.

In a similar way to Akedmir and Ciloglu[28], Yuan *et al*.[35] proposed the HMM segmentation system modifications based on modeling the phone boundaries with a special 1-state HMMs that are added to the HMMs phone models to improve segmentation accuracy[35].

Brognaux and Drugman[36] focused on a special case of Hidden Markov Model (HMM) based segmentation system in which the models are trained on the same corpus to align. The main advantage of this technique is that it does not require manually-aligned data and can be applied to any language. Brognaux studied first the impact of various training parameters (e.g., models configuration, number of training iterations) on the alignment accuracy with corpora varying in speaking style and language. Based on this study, Brognaux investigated the use of supplementary acoustic features and proposed two novel approaches: an initialization of the silence models based on a Voice Activity Detection (VAD) algorithm and the consideration of the forced alignment of the time reversed sound. Assessment was carried out on 12 corpora of different sizes, languages (some being under-resourced) and speaking styles. Experimental results show that the use of additional acoustic features increases the segmentation accuracy and that the use of VAD achieves very notable improvement, correcting > 60 % of the errors superior to 40 msec. Finally, combining the three improvement methods was also showed to provide the highest improvement with very low variability across the corpora, regardless of their size, improving the alignment rate by 8-10 % absolute.

PERFORMANCE COMPARISON

In Table 1, we provide a detailed performance comparison of the main segmentation methods exposed above, in terms of segmentation accuracy. We note that on TIMIT dataset Zhao *et al*.[20] achieves the best accuracy using the hybrid post processing techniques.

Table 1: Comparison of the main segmentation methods

References	Dataset	Features	Classifier	Accuracy
Toledano *et al*.[12]	Castilian Spanish: VESLIM corpus	MFCCs, Δ, ΔΔMFCCs,	HMMs, Context Dependent HMMs (CDHMMs), Context Independent HMMs (CIHMMs), Statistical Correction of Context Dependent Boundary Marks (SCCDBM)+Speaker Adaption (SA)+HMMs	Tolerance≤10 msec 87.18%
Adell *et al*.[32]	TALP Research Center corpus	MFCCs, Mel-Frequency Power Cepstrums (MFPC) Δ, ΔΔ MFPC, Δ Energy, Zero Crossing Rate (ZCR), mean frequency before and after boundary	HMMs, Artificial Neural Networks (ANNs), Regression Tree (RT), Dynamic Time Warping (DTW)	Tolerance a.≤10 msec: 82.00% b. ≤15 msec: 91.00%

Table 1: Continue

References	Dataset	Features	Classifier	Accuracy
Lee[26]	Korean TTS database	MFCCs, Δ, ΔΔ MFCCs,	HMMs, HMMs+Single Multilayer Perceptron (MLP), HMMs+Multiple MLPs, HMMs+Multiple MLPs (retraining)	Tolerance ≤20 msec: male: 93.2%, female: 93.9%
Park and Kim[19]	Korean TTS research database	MFCCs, Δ, ΔΔ MFCCs, Normalized log-energy	Context-independent HMMs, Context-dependent HMMs	Tolerance ≤20 msec: 97.05%
Jarifi *et al*.[23]	French corpus, English Corpus	MFCCs, Δ, ΔΔ MFCCs,	HMMs, GMM+HMMs, Brants Generalized Likelihood Ratio (GLR)	FR corpus: 10 msec: 79.90%, EN corpus: 10 msec: 81.71%
Hosom[7]	TIMIT database	Low-Energy Cepstral Mean Subtraction (LECMS+Δ)	HMM/ANN	Tolerance ≤10 msec: 79.47 Tolerance ≤20: 93.36
Mprose *et al*.[18]	TIMIT database	MFCCs, LFCCs, HFCC-E, PLP, WPF, SBC, MWP-ACE	Combination of multiple classifiers: LR, MPL NN, SVR, Model Trees M5 & HMMs	Tolerance ≤10 msec: 71.43%
Zhao *et al*.[20]	TIMIT database	MFCCs, Δ, ΔΔ MFCCs with cepstral mean and energy normalization	SVM/LDA	Tolerance ≤10 msec: 81.31 with SVM 79.91 with LDA
Brognaux and Drugman[36]	12 languages	Spectral features like MFCCs	Hidden Markov Models (HMMs)	French neutral corpus: 30 msec tolerance: 94.84%;
Frihia and Bahi[16]	Arab Phone: Arabic languages	MFCCs, Δ,ΔΔ MFCCs	Hidden Markov Models (HMMs)+SVM	Tolerance ≤20 msec: 85.88

CONCLUSION

In this study, we exposed an in-depth survey of the different Text-Dependent phonetic segmentation algorithms that exist in literature so far. These algorithms fall into two groups: post processing techniques that are based on refining the initial segmentation results through different procedures and acoustic model modification techniques.

Though modifications of the acoustic models may improve the accuracy of the segmentation[20], however, applying such kind of methods needs a restructuring and re-training of all the acoustic models and thus, they cannot by applied to existing segmentation systems. The post-processing techniques can be directly applied to any existing segmentation systems which make them more convenient and flexible than their counterparts. Also, some acoustic modeling techniques may sacrifice the segmentation results in a certain range to improve the overall performance as noted by Zhao *et al*.[20]. For example, the segmentation accuracy within 5 msec is reduced by Hosom[7] after applying the proposed model.

We have noted that with the exception of the research done by Mprose *et al*.[16], almost all the reviewed studies used MFCCs or MFCCs like as acoustic features, though wavelet based features has achieved better performance in phoneme recognition task[37].

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. 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.
17. 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.
18. Mporas, I., T. Ganchev and N. Fakotakis, 2010. Speech segmentation using regression fusion of boundary predictions. Comput. Speech Lang., 24: 273-288.
19. Park, S.S. and N.S. Kim, 2007. On using multiple models for automatic speech segmentation. IEEE. Trans. Audio Speech Lang. Process., 15: 2202-2212.
20. Zhao, S., Y. Soon, S.N. Koh and K.K. Luke, 2015. A hybrid refinement scheme for intra-and cross-corpora phonetic segmentation. Comput. Speech Lang., 29: 81-97.
21. Matousek, J., D. Tihelka and J. Psutka, 2003. Automatic segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction. Proceedings of 8th European Conference on Speech Communication and Technology, September 1-4, 2003, Eurospeech, Geneva, Switzerland, pp: 301-304.
22. Park, S.S. and N.S. Kim, 2006. Automatic speech segmentation based on boundary-type candidate selection. IEEE. Signal Process. Lett., 13: 640-643.
23. Jarifi, S., D. Pastor and O. Rosec, 2008. A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. Speech Commun., 50: 67-80.
24. Malfere, F., O. Deroo, T. Dutiot and C. Ris, 2003. Phonetic alignment: Speech synthesis-based vs. Viterbi-based. Speech Commun., 40: 503-515.
25. Lin, C.Y., K.T. Chen and J.S.R. Jang, 2005. A hybrid approach to automatic segmentation and labeling for Mandarin Chinese speech corpus. Proceedings of the 9th European Conference on Speech Communication and Technology, September 4-8, 2005, Interspeech, Lisbon, Portugal, pp: 1553-1556.
26. Lee, K.S., 2006. MLP-based phone boundary refining for a TTS database. IEEE. Trans. Audio Speech Lang. Process., 14: 981-989.
27. Lo, H.Y. and H.M. Wang, 2007. Phonetic boundary refinement using support vector machine. Proceedings of the 2007 IEEE International Conference on Acoustics Speech and Signal Processing, April 20-15, 2007, IEEE, Honolulu, USA., pp: 933-936.
28. Akdemir, E. and T. Ciloglu, 2010. HMM topology for boundary refinement in automatic speech segmentation. Electron. Let., 46: 1086-1087.
29. Lin, C.Y. and J.S.R. Jang, 2007. Automatic phonetic segmentation by score predictive model for the corpora of mandarin singing voices. IEEE. Trans. Audio Speech Lang. Process., 15: 2151-2159.
30. Mporas, I., T. Ganchev and N. Fakotakis, 2008. Phonetic segmentation using multiple speech features. Int. J. Speech Technol., 11: 73-85.
31. Stolcke, A., N. Ryant, V. Mitra, J. Yuan, W. Wang and M. Liberman, 2014. Highly accurate phonetic segmentation using boundary correction models and system fusion. Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-9, 2014, IEEE, Florence, Italy, pp: 5552-5556.
32. Adell, J., A. Bonafonte, J.A. Gomez and M.J. Castro, 2005. Comparative study of automatic phone segmentation methods for TTS. Proceedings of the IEEE International Conference Acoustic Speech Signal Processing (ICASSP'05), March 23-23, 2005, IEEE, Philadelphia, USA., pp: I/309-I/312.
33. Keshet, J., S. Shalev-Shwartz, Y. Singer and D. Chazan, 2007. A large margin algorithm for speech-to-phoneme and music-to-score alignment. IEEE. Trans. Audio Speech Lang. Process., 15: 2373-2382.

34. Kim, S., S. Yun and C.D. Yoo, 2011. Large margin discriminative semi-Markov model for phonetic recognition. IEEE. Trans. Audio Speech Lang. Process., 19: 1999-2012.
35. Brognaux, S. and T. Drugman, 2015. HMM-based speech segmentation: Improvements of fully automatic approaches. IEEE/ACM. Trans. Audio Speech Lang. Process., 24: 5-15.
36. 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.
37. Yuan, J., N. Ryant, M. Liberman, A. Stolcke, V. Mitra and W. Wang, 2013. Automatic phonetic segmentation using boundary models. Interspeech, 1: 2306-2310.