A comparative study of constrained and unconstrained approaches for segmentation of speech signal

Venkatesh Keri[†], *Kishore Prahallad*[†‡]

[†]International Institute of Information Technology, Hyderabad, India.
[‡]Language Technologies Institute, Carnegie Mellon University, USA.
venkateshk@research.iiit.ac.in, kishore@iiit.ac.in

Abstract

In this work, we compare different approaches for speech segmentation, of which some are constrained and the remaining are unconstrained by phone transcript. A high accuracy speech segmentation can be obtained by approaches constrained by phone transcript such as HMM forced-alignment when *exact phone transcript* is known. But such approaches have to adjust with *canonical phone transcript*, as *exact phone transcript* is tough to obtain. Our experiments on TIMIT corpus demonstrate that ANN and HMM phone-loop based unconstrained approaches, perform better than HMM forced-alignment based approach constrained by *canonical phone transcript*. Finally a detailed error analysis of these approaches is reported.

Index Terms: HMM, Group-Delay, ANN, Speech Segmentation.

1. Introduction

Speech segmentation is a process of identifying the boundaries between words, syllables or phones in a spoken utterance. Identification of boundaries at phone level is a difficult problem due to the phenomenon of co-articulation of speech sounds, where one sound may be modified in various ways by the adjacent sounds due to which the sounds may split or even disappear. This phenomenon may happen between adjacent words just as easily as within a single word [1] [2]. In the present work we provide a comparative study between constrained, and unconstrained approaches for segmentation which are defined based on the amount of knowledge used during speech segmentation.

Constrained segmentation is an approach where the knowledge of *number and sequence* of phones expected in the speech signal is used to obtain phone level boundaries. An example of such constrained approaches are Hidden Markov Model (HMM) Force-Alignment based segmentation [3], discriminatively trained Support Vector Machine (SVM) based segmentation [4] etc,. Such approaches have been shown to attain high performance using *exact phone transcript* as described in following works. Brugnara *et. al.* reported a boundary agreement of 88.8% with in 20 ms using HMMs[3], Joseph *et. al.* reported a boundary agreement of 92.3% with in 20 ms using discriminatively trained SVMs [4], Hosom *et. al.* reported a boundary agreement of 93.36% within 20 ms [5]. In practice, *exact phone transcript* is obtained manually by listening to the speech signal, which is very costly and tedious task. So, constrained segmentation approaches used a phone transcript generated from a canonical pronunciation dictionary which is termed as *canonical phone transcript*. For each word, a canonical pronunciation dictionary includes only the standard phone sequence assumed to be pronounced in read speech without any alternate pronunciations. In such *canonical phone transcript* insertions, deletions and substitutions of phones are unavoidable as observed in [6].

Unconstrained segmentation is an approach where the knowledge of number and sequence of phones expected in the speech signal is neither used nor assumed to obtain phone level boundaries. Such approaches focus on acoustic cues to detect the transient behavior at the phone boundaries. As these approaches do not use the phone sequence, boundary insertions and deletions are unavoidable as in [7][8][9]. These approaches can further be classified into heuristic based and model base unconstrained approaches. Heuristic based approaches mostly use some form of peak-picking algorithm to detect the boundaries and so does not require any training such as, Group Delay Function (GDF) based approach where Golipour *et. al.* reported 15.04% deletions and 6.6% insertions over the total number of manual boundaries in the corpus [7]; Perceptual Critical-band based approach where Aversano *et al.* reported 26% deletions and no insertions over the total number of manual boundaries in the corpus [8]. Model based approaches on the other hand requires some data and an algorithm to train the models. Suh *et. al.* [9] trained a frame-level boundary/non-boundary classifier on a Korean single speaker read speech database and reported a boundary agreement of 87% within 15 ms and 9% insertions over the total number of non-boundary frames in the database. Another model based unconstrained approach is to use the traditional phone HMM models in a phone-loop mode instead of forced-alignment mode which does not require phone transcript for segmentation.

In this paper, we have compared four approaches: HMM forced-alignment (constrained), HMM phone-loop (model unconstrained), GDF (heuristic unconstrained) and ANN (model unconstrained) based approaches for speech segmentation at phone level, and provide results and error analysis on TIMIT read speech database. Our results demonstrate that while model based unconstrained segmentation approaches performs better than constrained HMM forced-alignment approach using *canonical transcript*, the ANN based segmentation performs better than other unconstrained approaches.

This paper is organized as follows: Section 2 gives the brief description of four approaches. Evaluation criteria is described in section 3. Section 5 describes the experiments, results and analysis of the all the approaches.

2. Speech Segmentation Approaches

2.1. HMM-based Segmentation

The main advantage of using HMM models for speech segmentation is that it is built using the extensive knowledge and infrastructure of speech recognition. Just as in speech recognition, HMMs for speech segmentation are also trained using the stan-

dard EM algorithm. State sequence Θ is generated from *canonical phone transcript* and observation sequence O is obtained by parameterizing the speech signal. Speech parametrization is performed by computing a feature vector every 5 ms using a 10 ms Hamming window and a pre-emphasis coefficient of 0.97. The feature vector used for HMM-based segmentation is a 12 Mel-Frequency Cepstral Coefficients (MFCCs) with Cepstral Mean Normalization (CMN) and normalized log energy, as well as their first and second order differences yielding a total of 39 components. To compute the likelihood function, state sequence Θ is considered to be hidden data. Thus in order to obtain a maximum likelihood estimate $\bar{\lambda}$ of the model parameters, we must calculate the conditional expectation of the likelihood given a current set of parameters λ. Objective function $Q(\lambda, \bar{\lambda})$ has to be maximized in successive iterations:

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} P(O, \theta|\lambda) \log P(O, \theta|\bar{\lambda}) \quad (1)$$

Even though both speech recognition and speech segmentation use HMMs, however, it is important to realize that the goals of both the tasks are different. Hence these differences are reflected in the topology of HMM models. Past research [10][11] have indicated that context-independent models are preferred over context-dependent models and almost no improvement beyond two gaussians per state for speech segmentation task. HMM topology of each phone is context-independent, 5 state sequential, left-to-right without any skip-state and observation probability distribution of each state is characterized by 2 mixture gaussians. A speaker-independent HMM models are trained and tested using HMM toolkit [12]. Following are the two approaches for speech segmentation using HMMs:

2.1.1. Forced-Alignment (HMM-FA)

This is a constrained model-based approach, constrained by phone transcript. So, canonical phone transcript of the speech signal is required for segmentation. The corresponding phone HMM models are force-aligned with the parametrized observation sequence of speech signal to compute $[\log(P(O|\theta_t)]$ for every t and Viterbi search is employed to obtain the optimal segment boundaries.

2.1.2. Phone-Loop (HMM-PL)

This is an unconstrained model-based approach and hence does not require phone transcript for segmentation. As phone transcript is not given, it assumes that all the phone are equally likely for whole of the observation sequence. Log likelihood of all state for each observation vector is computed as $[\log \max_{t=1}^{T} P(O|\theta_t)]$ (where T is total number of phone models) and Viterbi search is employed to obtain the optimal segment boundaries.

2.2. GDF-based Segmentation

GDF-based segmentation is an example of heuristic based unconstrained approach which focus on acoustic cues to detect the transient behavior at the phone boundaries. Brief description of this work is presented here and more details can be found in [7].

Speech signal is parametrized into overlapping frames of 8 ms frame size and 4 ms frame shift. A smoothed power spectrum $S(w, n)$ is computed by applying a 4-by-4 median filter on a 512 point FFT spectrogram $X(w, n)$. Compute the gradient of $S(w, n)$ to obtain a measure for the change in the energy at different frequency in the speech signal. These energy changes are summed over in 5 different bands i.e., 0-8000Hz, 0-500Hz, 500-1420Hz, 1420-2386Hz and 2386-8000Hz to obtain different $Y(n)$ for each band. Now as this can be posed as a peak-picking problem, a modified group-delay function is applied on each $Y(n)$ separately to obtain boundaries using equation (2). An "OR" operation is performed on the boundaries obtained by different bands to obtain final boundaries.

$$\tau_{Y(n)} = sign \left| \frac{Y_R(n)Z_R(n) + Y_I(n)Z_I(n)}{S(n)^2} \right|^{\alpha} \quad (2)$$

2.3. ANN-based Segmentation

In order to perform speech segmentation, we employed Artificial Neural Networks (ANN) as a classifier in this work. Artificial Neural Networks consists of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between the two nodes has a weight associated with it [13]. As ANN models are known for their classification abilities in nonlinear space, we have used a multi-layer feed-forward neural network to build a boundary / non-boundary classifier. The input to such a classifier is acoustic features extracted at frame t from the speech signal, and the ANN model is expected to classify it as boundary or non-boundary frame [9].

2.3.1. Training an ANN Model

In order to train an ANN model, class labels (boundary / non-boundary) for each of the feature vectors are required. Class labels are obtained from the manually marked boundaries in the train-set. Using manual boundaries, all the feature vectors at the boundary frames are labeled as examples of boundary class. As all the frames between any two adjacent boundaries are non-boundary frames, there will be huge imbalance between the number of feature vectors of each class in the training data, which could bias the classifier. Hence a frame which is in the middle of the two adjacent boundaries is selected as an example of non-boundary class.

Let $\boldsymbol{x}_t$ denote the feature vector extracted at frame t, then the input to the ANN model is an augmented feature vector $\hat{\boldsymbol{x}}_t = [\boldsymbol{x}_{t-l}, .., \boldsymbol{x}_t, .., \boldsymbol{x}_{t+l}]$. The value of l denotes the number of neighboring feature vectors appended to $\boldsymbol{x}_t$. Given $\hat{\boldsymbol{x}}_t$, the corresponding class label $\boldsymbol{y}_t = [a_t \ b_t]$ is created, where a_t and b_t are boundary and non-boundary evidence respectively for frame t whose values depend on the output target function used for training the network. With these input and output data, an ANN classifier is trained using back propagation learning to adjust the weights of neural network so as to minimize the mean squared error between the actual and the desired output.

2.3.2. Segmentation using ANN Model

Steps for speech segmentation: *Acoustic Score:* Feature vectors are extracted from the speech signal for each frame t as described in Section 2.3.1. The corresponding augmented vector $\hat{\boldsymbol{x}}_t$ is given to the ANN classifier, to obtain boundary/non-boundary evidences. If $\hat{a_t}$ and $\hat{b_t}$ are the obtained boundary and non-boundary evidences respectively, then acoustic score of $\hat{\boldsymbol{x}}_t$ is computed as $A(\hat{\boldsymbol{x}}_t) = \hat{a_t} - \hat{b_t}$. In order to remove some spurious peaks in A, a 5-point linearly weighted mean smoothing is applied to obtain $\hat{A}$.

Detection of Boundary Region: For each frame t, if $\hat{A}(\hat{\boldsymbol{x}}_t) > 0$, then it is classified as boundary else non-boundary. The region of consecutive boundary frames without any non-

boundary frame is considered as a *boundary region*. Let $\imath_q$ and $\jmath_q$ denote the begin and end frames of a boundary region q.

Location of Boundary: Once the boundary regions are detected, exact location of boundary in each of these regions has to be located. Among the frames in each region q, the frame with highest acoustic score is marked as boundary frame using equation (3), where $\hat{i}_q$ is the index of the boundary frame from region q.

$$\{\hat{i_q}\} = \arg\max_t \{\hat{A}(\hat{\boldsymbol{x}}_t)\}_{t=\imath_q}^{\jmath_q} \quad (3)$$

In this paper, l is taken as 5, so each feature vector is a concatenation of 11 frames. Total dimension for each input feature vector is 143 (11 frames x 13 coefficients). As output target function used for training an ANN model is tangential (N), output vector $\boldsymbol{y}_t$ is [1 -1] and [-1 1] for boundary and non-boundary frames respectively.

3. Evaluation Criteria

The performance of speech segmentation is evaluated using the following five metrics, essentially by comparing the predicted boundary with the manually marked boundary in the speech signal. If ζ_i denote the time stamp of the manually marked boundary i in the speech signal, then a region of tolerance ϵ_i is defined as $(\zeta_i - (\zeta_i - \zeta_{i-1})/2) \leq \epsilon_i \leq (\zeta_i + (\zeta_{i+1} - \zeta_i)/2)$. For every i, if there exists a predicted boundary $\hat{i}$ with its time stamp denoted by $\zeta_{\hat{i}}$, such that $\zeta_{\hat{i}}$ is within ϵ_i, then $\hat{i}$ is considered as correct boundary. If there are more than one predicted boundary within ϵ_i then one of the predicted boundaries which is nearest to ζ_i is considered as correct boundary, and the rest are considered as inserted boundaries. If there is no predicted boundary within ϵ_i, then i is considered as deleted boundary.

RMS Error: It is the root mean square of the deviations between the manual and its nearest correct boundaries.

Agreement Percentage (AGR): It is the percentage of correct boundaries with a tolerance (absolute deviation) of less than τ ms over the total number of correct boundaries.

Boundary Error Rate (BER): It is defined as the summation of insertion (INS) and deletion (DEL) percentages. Here, the INS percentage is computed as number of insertions over the total number of manual boundaries, and the DEL percentage is computed as number of deletions over the total number of manual boundaries.

Performance of the segmentation is better when RMS, DEL, INS & BER are low and AGR is high.

4. Results and Discussion

All our experiments are conducted on TIMIT corpus, which is recorded in a clean environment at 16 kHz sampling rate and has been labeled manually using 61 phones. Excluding "SA" files, this corpus has 3696 training files and 1344 testing files [5]. Before analyzing the speech segmentation performance, we have to analyze the similarity between the *exact* and *canonical* phone transcripts and errors in *canonical* phone transcript. This can be analyzed by aligning both the transcripts using simple dynamic programing and computing the number phone insertions, deletions and substitutions by *canonical* phone transcript over *exact* phone transcript Table 1 shows that 83.7% of the *canonical* phone transcript matches with *exact* phone transcript and former has a PER of 27.7% over the latter. Observation probability distributions of HMM states are experimented using 1, 2 and 8 mixture gaussians and found that 2 GMMs' performed slightly better than 1 and 8 GMMs' both for HMM-FA and HMM-PL which is in agreement with [10][11]. In order to check the usability of these three approaches on non-native English and other languages other than US English (TIMIT-test) using the existing models, two more test-sets are created. First, an Indian English test-set (INE) of 10 TIMIT test prompts are recorded by two speaker with Indian English accent. Secondly, a different language i.e., Telugu test-set (TEL) of 20 Telugu sentences spoken by native Telugu speaker. All the four approaches are tested on TIMIT-test and INE as they are still English, but only HMM-PL, GDF & ANN based approaches can be tested on TEL, English phone set cannot be used to generate Telugu phone transcript. All the test-sets are manually labeled to compare the performances of all the approaches. Table 2 shows that ANN models trained on TIMIT database outperformed other approaches on all test-sets. Table 2 shows that model based unconstrained approaches trained on TIMIT-train database outperformed other approaches on all test-sets and among unconstrained approaches, ANN outperforms others. Another inference from this table is that HMM-PL performed better than HMM-FA with *canonical* phone transcript.

Table 1: *Accuracy, Phone Error Rate, Substitutions, Insertions, and Deletions of phones by canonical phone transcript over exact phone transcript using TIMIT corpus.*

Data	Acc.	PER	Sub.	Ins.	Del.
Train	83.7%	27.7%	19146	16225	4205
Test	83.9%	27.1%	6685	5625	1612

Table 2: *Performance of all the approaches on US English (TIMIT-test), Indian English (INE) & Telugu (TEL) Test-sets.*

Approach	Test-Set	RMS (ms)	AGR % ($\tau \leq 20ms$)	DEL (%)	INS (%)	BER (%)
HMM-FA	TIMIT-test	15.2	82.53	10.75	19.97	30.72
	INE	14.1	85.58	12.29	18.19	30.48
HMM-PL	TIMIT-test	15.7	81.71	17.33	9.75	27.08
	INE	12.5	84.08	16.63	11.33	27.96
	TEL	15	82.11	15.81	20.86	36.67
GDF	TIMIT-test	11.4	88.56	24.51	12.3	36.81
	INE	9.3	91.48	19.64	15.30	34.94
	TEL	8.8	93.02	31.51	13.75	45.26
ANN	TIMIT-test	9.3	92.18	13.91	7.81	21.72
	INE	9.2	91.85	17.47	7.59	25.06
	TEL	11.6	93.84	21.52	11.59	33.11

Table 1 shows that, apart from substitutions, most of the errors are caused by insertions in canonical phone transcript. This is directly reflected on the performance of HMM-FA with high INS and hence high BER as shown in table 2. As HMM-PL does not use phone transcript for segmentation, INS is lower and hence lower BER as shown in table 2. Even though HMM-PL does not use transcript for segmentation, it uses canonical phone transcript for training HMM's. On the other hand ANN is not constrained by phone transcript neither for training nor for segmentation and hence outperforms other approaches.

Table 3 shows that ANN based approach not only out performed GDF, HMM-PL and HMM-FA based approaches, but also performs as good as constrained approaches using *exact* phone transcript [3][4] except for $\tau \leq 10ms$.

In order to investigate the different types of boundary errors caused by each of these approaches, a detailed analy-

Table 3: *Agreement Percentage of correctly predicted boundaries by some previous works, HMM-FA, HMM-PL, GDF and ANN based approaches for different tolerance (τ) values.*

	Approach	AGR % with $\tau \leq$			
		$10ms$	$20ms$	$30ms$	$40ms$
Using exact phone trans.	Brugnara *et. al.*[3]	74.6	88.8	94.1	96.8
	Joseph *et. al.*[4]	80.0	92.3	96.4	98.2
	Hosom *et. al.*[5]	79.30	93.36	96.74	98.22
Using canonical phone trans.	HMM-FA	55.85	82.51	94.76	98.19
Without using phone trans.	HMM-PL	51.82	81.71	94.89	98.16
	GDF	42.97	88.56	96.31	98.27
	ANN	59.10	92.18	97.39	99.06

sis of boundary deletion and insertion is performed. All the phones are grouped into five broad phonetic classes i.e., fricatives, nasals, stops (closure + burst), semi-vowels, vowels. Table 4 shows boundary deletion percentage of each class pair (CPDEL) and boundary insertion percentage of each class (CINS) for all four approaches:

$$CPDEL = \frac{\#\ Deleted\ Class\ Pairs}{\#\ Class\ Pairs}\ X\ 100 \quad (4)$$

$$CINS = \frac{\#\ Insertionsin\ Class}{\#\ Class\ instances}\ X\ 100 \quad (5)$$

A smaller value of CPDEL, CINS indicates a better segmentation performance. From this table, we can infer that CPDEL of ANN segmentation is least for boundaries between VOW/SVOW and FRI/NAS/STOP; STOP and VOW; FRI and STOP. HMM-FA performed better than ANN for rest of the class pairs. Another inference is that number of insertions in all the class is least for ANN based approach.

5. Conclusion

In this paper, we have compared the constrained (HMM-FA) approach with *canonical* phone transcript and unconstrained (HMM-PL, GDF and ANN) approaches for segmentation of speech signal. Our results demonstrate that while model based unconstrained approaches perform better than constrained approaches using *canonical* phone transcript, the ANN based segmentation out performs other approaches. HMM-FA performs poorly in comparison with HMM-PL and ANN could be justified from the large number of consonants in TIMIT corpus which are incomplete due to missing closures or release [2]. We have also shown that the ANN based models trained on TIMIT database could be used to segment non-native English and Telugu speech data.

6. References

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, united states ed ed. Prentice Hall PTR, April 1993.

[2] P. K. Ghosh and S. S. Narayanan, "Closure duration analysis of incomplete stop consonants due to stop-stop interaction," *Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. EL1–EL7, Jul. 2009.

[3] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models," *Speech Commun.*, vol. 12, no. 4, pp. 357–370, 1993.

[4] J. Keshet and S. Shalev-shwartz, "Phoneme alignment based on discriminative learning abstract," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005.

Table 4: *Deletion percentages of each class pair and insertion percentages of each class are computed for HMM-FA, HMM-PL, GDF & ANN based approaches. All the cells in the table where ANN is performing better than others are shaded.*

Class		Deletions					Insertions
		FRI	NAS	STOP	SVOW	VOW	
FRI	# Pairs	406	98	1388	471	4615	6978
	HMM-FA	17.00	6.12	6.48	2.76	3.03	15.92
	HMM-PL	38.67	6.12	15.42	3.40	6.02	7.14
	GDF	35.96	9.18	5.26	9.98	11.18	15.87
	ANN	38.67	11.22	5.55	3.18	4.25	7.09
NAS	# Pairs	675	72	1062	288	1804	3901
	HMM-FA	8.74	31.94	14.41	11.81	4.27	20.05
	HMM-PL	15.26	77.78	37.48	20.14	12.97	6.05
	GDF	16.89	86.11	45.67	63.19	26.00	3.61
	ANN	15.26	88.89	33.62	43.40	8.20	4.56
STOP	# Pairs	1476	119	5436	1377	5345	13753
	HMM-FA	14.97	10.08	6.42	6.61	13.53	18.53
	HMM-PL	23.85	31.09	13.12	10.75	17.06	9.28
	GDF	23.17	41.18	20.03	17.28	24.81	8.47
	ANN	24.25	36.13	10.89	10.24	13.43	5.57
SVOW	# Pairs	304	120	462	97	3915	4898
	HMM-FA	22.70	10.00	19.48	23.71	16.53	14.66
	HMM-PL	21.05	19.17	26.41	28.87	38.11	4.61
	GDF	6.58	24.17	8.23	36.08	55.33	9.51
	ANN	2.30	5.83	3.25	26.80	33.97	6.37
VOW	# Pairs	4271	3499	5225	2444	1286	16725
	HMM-FA	7.38	12.37	10.16	27.66	17.26	18.69
	HMM-PL	10.09	16.18	14.58	40.75	32.58	10.34
	GDF	6.77	27.92	6.56	60.35	60.34	14.82
	ANN	2.55	7.40	3.25	48.77	48.52	8.29

[5] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Commun.*, vol. 51, no. 4, pp. 352–368, 2009.

[6] A. Vorstermans, J.-P. Martens, and B. Van Coile, "Automatic segmentation and labelling of multi-lingual speech data," *Speech Commun.*, vol. 19, no. 4, pp. 271–293, 1996.

[7] G. Ladan and D. O'Shaughnessy, "A new approach for phoneme segmentation of speech signals," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[8] G. Aversano, A. Esposito, and M. Marinaro, "A new text-independent method for phoneme segmentation," in *Proceedings of the 44th IEEE Midwest Symposium on Circuits and Systems*, Orlando, Florida, 2001.

[9] 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, 3-6 1996, pp. 1297 –1300 vol.3.

[10] D. Toledano, L. Gomez, and L. Grande, "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 617 – 625, nov. 2003.

[11] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Comput. Speech Lang.*, vol. 24, no. 2, pp. 273–288, 2010.

[12] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen, *The HTK Book for HTK V2.0*. Cambridge University Press, Cambridge, UK, 1995.

[13] B. Yegnanarayana, *Artificial neural networks*. Prentice Hall of India, 2004.