QUALITATIVE EVALUATION AND ERROR ANALYSIS OF PHONETIC SEGMENTATION

Arif Khan[1−3], *Ingmar Steiner*[1,2]

[1]*Multimodal Computing and Interaction, Saarland University, Germany*
[2]*German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany*
[3]*Saarbrücken Graduate School of Computer Science, Germany*
arifkhan@coli.uni-saarland.de

Abstract: Speech segmentation is the process of splitting and identifying the boundaries between different units of speech, i.e., words, syllables, and phones. This paper focuses on the automatic phonetic segmentation of speech and the methods used for its evaluation. We explain the current methods used for the evaluation of speech segmentation and highlight the details that have not been sufficiently addressed in the literature. Several metrics are explained for analysis. The phones are grouped into several classes and the phone class transitions are observed. We found that, most of the errors comes from those class transitions which are also difficult for humans to segment.

1 Introduction

Nowadays many speech based applications are used in our daily life. The correctness of these applications directly depends on good training data [1]. A good training data for such systems is one, which defines accurate boundary location (time stamps) along with its labels, more formally the segmentation of speech. Speech can be segmented on several different levels, e.g., words, syllables, phones. In this paper, we will discuss the phone level segmentation i.e., splitting the speech into distinct phones.

The naive way to segment speech is to do it manually, also called manual segmentation. In this method, the phoneticians places the boundaries by hand using a computer program e.g., Praat [2]. This method produces the most accurate segmentation, however, it is time consuming and is prone to individual errors [3]. Therefore, automatic segmentation is desired which can segment the speech easily and with little effort, compared to manual segmentation. Moreover, this method is conveniently reproducible and the exact segmentation is obtained every time it is applied.

Many researchers have used hidden Markov models (HMMs) for automatic segmentation [1, 4, 5]. A generic approach in their work was to force align the speech with their corresponding transcripts in a supervised fashion. Some researchers have also shown how the different variants of HMMs, i.e., monophones or triphones, affects the segmentation quality [6]. Researchers have also used different modifications to the HMM based technique to improve the performance. Zhao et al. [7] used statistical corrections of the segmented boundaries along with different step sizes for feature extraction to improve the segmentation performance.

The evaluation of automatic segmentation is carried out by comparing it against a reference segmentation. The reference segmentation could be a manual segmentation or a segmentation generated by another tool. But for evaluation, the manual segmentation is considered to be the best reference [8]. Usually, the accuracy (Section 2.1) of the segmented speech is computed. While the accuracy shows the quality of the segmented speech, in most of the literature it is

not clear how it is computed. This poses a potential problem, because the results vary greatly depending on the way accuracy is computed. Räsänen et al. [9] highlighted this problem and mentioned a few factors that influence the accuracy. One such problem is related to the selection of a threshold for comparing boundaries. The segmentation accuracy improves as we select a higher threshold. This is because – as the search space around the boundary is increased (by taking a higher threshold) – the likelihood of a segmented boundary to be found also increases considerably and as a result, a higher accuracy is achieved. Another problem is that the same segmented boundary can be considered for two different reference boundaries. Specifically, this problem increases with higher thresholds. Therefore, only reporting the accuracy is not sufficient and the way it is computed should also be explained. In this paper, we mention in detail the steps to find the accuracy in Section 2.1 and show how it varies with various threshold.

The rest of the paper is organized as follows. In the next section, we discuss several methods that can be used for the evaluation of speech segmentation. In Section 3, we explain our experiments in detail. This includes the speech corpora and the phone set that is used. In Section 4, the results of our experiments are given and discussed. Finally, the conclusion and future work are given in Section 5.

2 Methods for Analysis

The segmentation accuracy is directly related to the phone recognition accuracy [1], i.e., high phone recognition gives better segmentation accuracy. Therefore, we checked the overall phone accuracy of the segmented speech in two ways. First, by looking at the phone and word error rates and then by comparing the distribution of phones in the reference and segmented phone sets. The phone error rate (PER) and word error rate (WER) are computed through the Levenshtein distance (LD), which is given as:

$$LD = \frac{I + D + S}{N} * 100 \tag{1}$$

Where I, D, and S are the number of insertions, deletions, and substitutions respectively, and N is the total number of phones or words in the reference. The results are given in Section 4.1.

2.1 Accuracy

The direct method of evaluating the segmentation is to measure the accuracy, i.e., to find the number of correct boundaries. The accuracy is usually given as a percentage and is calculated as:

$$Accuracy = \frac{CorrectBoundaries}{TotalBoundaries} \times 100 \tag{2}$$

Based on [10], the following method was used for finding the accuracy.

2.1.1 Boundary Comparison Method

For the given sets of reference (**ref**) and segmented (**seg**) phone boundaries, we proceed as follows. First, make a search space of 40 ms (20 ms to the left and 20 ms to the right) around each reference boundary. If the search spaces of two **ref** boundaries overlap then *shrink* the search spaces by truncating at the middle of the overlapping area. This shrinking of the search space is done to prevent a single **seg** boundary from appearing in the search space of two neighbouring **ref** boundaries.

For comparison, a single boundary from the **ref** set is taken along with its search space. Any **seg** boundary that lies within the search space of this **ref** boundary is considered a match,

Set	# Speakers	# Sentences	# Hours
Training	462	3969	3.14
Core test	24	192	0.16
Complete test set	168	1344	0.81

Table 1 – Details of the TIMIT corpus

otherwise a miss (deletion). If more than one boundary is found, then the extra boundary is considered an insertion. This is repeated for all **ref** boundaries. The final result is obtained by taking the mean of the results from all the utterances.

3 Experiment Setup

3.1 Software

We used HMMs [11, 12] to train the acoustic models and obtain the segmentation. 12 mel-frequency cepstral coefficients (MFCCs) along with their first and second order derivatives and normalized log energy are used as acoustic features. Altogether they form a composite feature vector of 39 dimensions. The Montreal Forced Aligner (MFA) framework [13] is used to produce the segmentation. The MFA is a python wrapper around the Kaldi [14] speech recognition toolkit. The MFA can be used in two ways, i.e., either to train an acoustic model from a corpus and then use that model for segmentation, or to use a pre-trained model and perform the segmentation directly with it. We use the first approach.

3.2 Data

We used the TIMIT [15] corpus for our experiments. It consists of phonetically balanced recordings of prompted English speech, recorded at 16 kHz with 16 bits per sample. TIMIT contains a total of 6300 sentences (5.40 h) spoken by 630 speakers with 10 sentences per speaker. The corpus comes with separate train and test sets as shown in Table 1. All sentences are manually segmented at the phone level.

The corpus is grouped into eight different American English dialects. There are three types of sentences in the TIMIT corpus. The *(SA)* sentences, which are dialect sentences and show the dialects of the speakers; *(SI)* sentences, which are phonetically diverse sentences; and *(SX)* sentences, which are phonetically compact sentences. The *(SA)* sentences consist of the same word sequences (text) in the train and test sets. We used the complete train set for training and the complete test set (except *SA*) for evaluation.

3.3 Phone Set

The TIMIT corpus manual segmentation uses a set of 61 phones, shown in Table 2.

However, the majority of previous work on the TIMIT corpus uses a smaller phone set, either 48 or 39. Therefore, the 61 phone set is reduced to a smaller phone set as follows. The *pau*, *epi*, *#h* phones are mapped to *sil*. The *q* phone is dropped. The phones can have different behavior when they are represented as stressed or unstressed (usually by appending a 0, 1, or 2 to a vocalic phone). As the manual segmentation of TIMIT has no stress markings, we therefore also use phones without stress markings to facilitate comparison with the manual segmentation.

	Phone Label	Example		Phone Label	Example		Phone Label	Example
1	iy	b*ee*t	21	jh	*j*oke	41	em	botto*m*
2	ih	b*i*t	22	ch	*ch*oke	42	nx	wi*nn*er
3	eh	b*e*t	23	b	*b*ee	43	en	butto*n*
4	ey	b*ai*t	24	d	*d*ay	44	eng	Wash*ing*ton
5	ae	b*a*t	25	g	*g*ay	45	l	*l*ay
6	aa	b*o*tt	26	p	*p*ea	46	r	*r*ay
7	aw	b*ou*t	27	t	*t*ea	47	w	*w*ay
8	ay	b*i*te	28	k	*k*ey	48	y	*y*acht
9	ah	b*u*t	29	dx	mu*dd*y	9	hh	*h*ay
10	ao	b*ough*t	30	s	*s*ea	50	hv	a*h*ead
11	oy	b*oy*	31	sh	*sh*e	51	el	bott*le*
12	ow	b*oa*t	32	z	*z*one	52	bcl	*b closure*
13	uh	b*oo*k	33	zh	a*zu*re	53	dcl	*d closure*
14	uw	b*oo*t	34	f	*f*in	54	gcl	*g closure*
15	ux	t*oo*t	35	th	*th*in	55	pcl	*p closure*
16	er	b*ir*d	36	v	*v*an	56	tcl	*t closure*
17	ax	*a*bout	37	dh	*th*en	57	kc]	*k closure*
18	ix	deb*i*t	38	m	*mom*	58	q	*glottal stop*
19	axr	butt*er*	39	n	*noon*	59	pau	*pause*
20	ax-h	s*u*spect	40	ng	si*ng*	60	epi	*epenthetic silence*
						61	h#	begin/end marker

Table 2 – The TIMIT original phone set

4 Analysis

4.1 Phone Distribution and Levenshtein Distance

The distributions of phones in the reference and segmented sets is plotted in Figure 1. The x axis represents the phones and the y axis, the phone count in log scale. From the comparison, we can see that the distributions of phones in both sets is almost the same and that the segmentation has no major errors. The PER is 27.48 % with 6193 insertions, 545 deletions and 6560 substitutions and the WER is 1.62 % with 152 insertions, 1 deletion and 2 substitutions. From the PER and WER one can infer that the phone recognition and word recognition is high and segmentation is reasonably correct. After this step, the detailed analysis can be performed to see the accuracy of segmentation.

4.2 Segmentation Accuracy

We applied the method explained in Section 2.1.1 for phone boundary analysis. The boundary accuracies are shown in Figure 2. As one can see, for a threshold of 10 ms and 20 ms, an accuracy of 45 % and 82 % was obtained, respectively. The accuracy reaches to 96 % for a threshold of 50 ms.

4.3 Phone Transition Analysis

For phone transition analysis, the phones were grouped into seven different phonetic classes. An additional SIL class is also included, which represents the SIL (silence) phone. In Table 3, the different phones are shown along with their phone class and symbol. The missing boundary in the reference and its right context are converted into the corresponding phonetic class. For example, if a phone AO boundary was missing in the reference utterance, and the phone EL is

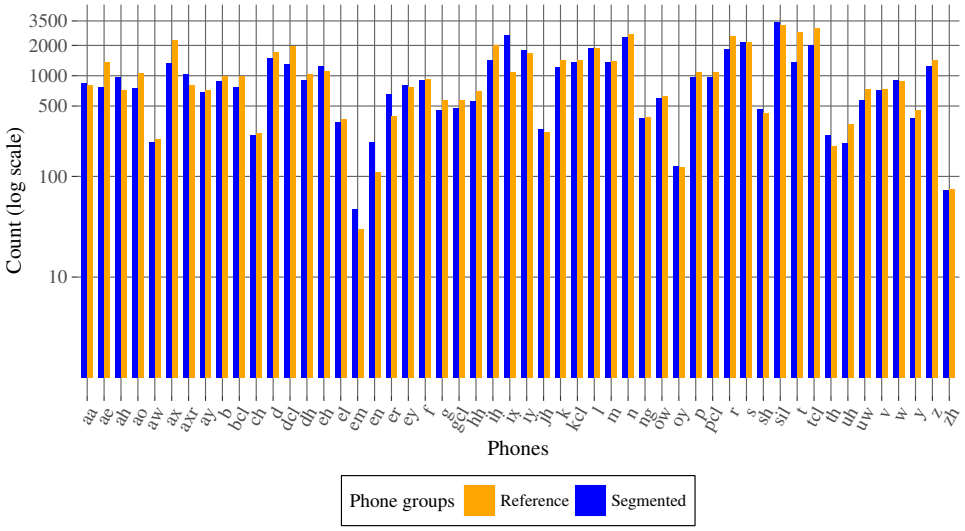

Figure 1 – Comparison of the phone distribution in the reference and segmented phone set.

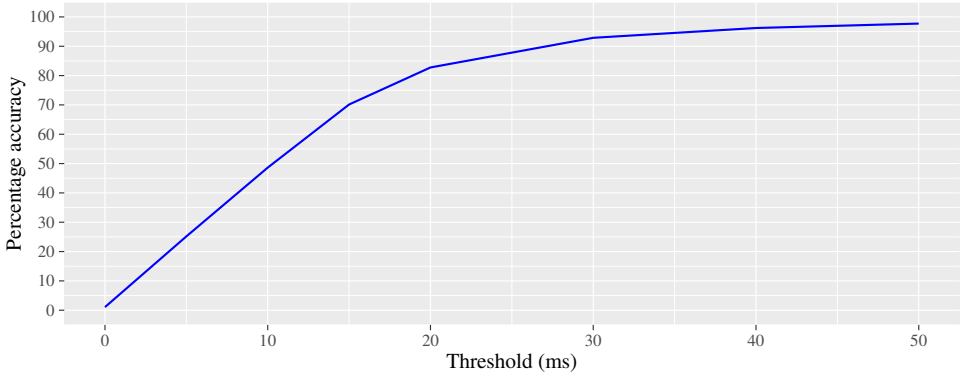

Figure 2 – Segmentation accuracy with varying thresholds

No.	Phone	Class	Symbol
1	AA, AE, AH, AO, AW, AX, AXR, AY, EH, ER, EY, IH, IX, IY, OW, OY, UH, UW	vowel	V
2	EL, HH, L, R, W, Y	semivowel-and-glides	G
3	EM, EN, M, N, NG	nasal	N
4	B, D, G, K, P, T	stop	S
5	BCL, DCL, GCL, KCL, PCL, TCL	unvoiced-stop	US
6	DH, F, S, SH, TH, V, Z, ZH	voiced-fricative	VF
7	CH, JH	unvoiced-fricative	UF
8	SIL	silence	SIL

Table 3 – Phone classes for analysing the segmentation

Figure 3 – Phone errors grouped by phone class transitions

in its right context, then we have a transition from class V to class G, as phone AO belongs to the vowel class and phone EL belongs to the semivowels-and-glides class.

We summed up all the erroneous class transitions and found a total of 8532 missing boundaries. As can be seen in Figure 3, the highest error class transitions are semivowel-and-glides to vowel and vowel to semivowel-and-glides with 23 % and 18 % errors respectively. The class unvoiced-stops to stops also contributes 12 % errors.

5 Conclusion and Future Work

In this paper, we have mentioned a few methods that can be used for the evaluation of segmentation. The quality of segmentation can be judged by computing the accuracy; however, we need to mention the details of how it was computed. The class based comparison shows the details of the errors and the most difficult phone transitions for segmentation. These class transitions are also the most difficult in manual segmentation because of the ambiguity in the boundary placement. In fact, Wesenick and Kipp [8] have stated that the boundaries which are difficult for humans to segment are also difficult for automatic systems.

The current approaches only consider the time stamps of the boundaries for deciding if a reference and segmented boundary is correct. Oftentimes, there are cases when the labels of the corresponding boundaries does not match, but because the time difference between them is less than a certain threshold, the boundaries are considered correct. For future work, we think a more sophisticated metric is required for boundary comparison. Such a metric should compare the label and context as well as the timestamps, to establish the alignment between the reference and segmented boundaries.

References

[1] TOLEDANO, D. T., L. A. H. GÓMEZ, and L. V. GRANDE: *Automatic phonetic segmentation. IEEE Transactions on Speech and Audio Processing*, 11(6), pp. 617–625, 2003. doi:10.1109/TSA.2003.813579.

[2] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer (version 6.0.23)*. 2017. URL http://www.praat.org.

[3] SVENDSEN, T. and F. SOONG: *On the automatic segmentation of speech signals*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 77–80. 1987. doi:10.1109/icassp.1987.1169628.

[4] LJOLJE, A., J. HIRSCHBERG, and J. P. VAN SANTEN: *Automatic speech segmentation for concatenative inventory selection*. In J. P. H. VAN SANTEN, R. W. SPROAT, J. P. OLIVE, and J. HIRSCHBERG (eds.), *Progress in Speech Synthesis*, pp. 305–311. Springer, 1997.

[5] WIGHTMAN, C. W. and D. T. TALKIN: *The Aligner: Text-to-speech alignment using Markov models*. In J. P. H. VAN SANTEN, R. W. SPROAT, J. P. OLIVE, and J. HIRSCHBERG (eds.), *Progress in Speech Synthesis*, pp. 313–323. Springer, 1997.

[6] BROGNAUX, S. and T. DRUGMAN: *HMM-based speech segmentation: Improvements of fully automatic approaches*. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(1), pp. 5–15, 2016. doi:10.1109/TASLP.2015.2456421.

[7] ZHAO, S., Y. SOON, S. N. KOH, and K. K. LUKE: *Phonetic segmentation using statistical correction and multi-resolution fusion*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6694–6698. 2013. doi:10.1109/ICASSP.2013.6638957.

[8] WESENICK, M.-B. and A. KIPP: *Estimating the quality of phonetic transcriptions and segmentations of speech signals*. In *International Conference on Spoken Language Processing (ICSLP)*, vol. 1, pp. 129–132. 1996. doi:10.1109/icslp.1996.607054.

[9] RÄSÄNEN, O. J., U. K. LAINE, and T. ALTOSAAR: *An improved speech segmentation quality measure: the R-value.* In *Interspeech*, pp. 1851–1854. 2009. URL http://isca-speech.org/archive/interspeech_2009/i09_1851.html.

[10] RÄSÄNEN, O. J.: *Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture*. Master's thesis, Helsinki University of Technology, 2007.

[11] RABINER, L. R.: *A tutorial on hidden Markov models and selected applications in speech recognition*. *Proceedings of the IEEE*, 77(2), pp. 257–286, 1989. doi:10.1109/5.18626.

[12] JUANG, B. H. and L. R. RABINER: *Hidden Markov models for speech recognition*. *Technometrics*, 33(3), pp. 251–272, 1991. doi:10.1080/00401706.1991.10484833.

[13] MCAULIFFE MICHAEL, S. M., MICHAELA SOCOLOF, M. WAGNER, and M. SONDEREGGER: *Montreal Forced Aligner*. 2017. URL http://montrealcorpustools.github.io/Montreal-Forced-Aligner.

[14] POVEY, D., A. GHOSHAL, G. BOULIANNE, L. BURGET, O. GLEMBEK, N. GOEL, M. HANNEMANN, P. MOTLÍČEK, Y. QIAN, P. SCHWARZ, J. SILOVSKÝ, G. STEMMER, and K. VESELÝ: *The Kaldi speech recognition toolkit*. In *IEEE Workshop on Automatic Speech Recognition and Understanding*. 2011.

[15] GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, D. S. PALLETT, N. L. DAHLGREN, and V. ZUE: *TIMIT acoustic-phonetic continuous speech corpus*. 1993.