SYLLABLE NUCLEUS DURATIONS ESTIMATION USING LINEAR REGRESSION BASED ENSEMBLE MODEL

Jingli Lu[1,2]*, Ruili Wang*[1,2]*, Liyanage C De Silva*[1,3] *, Yang Gao* [2,4]

[1]School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand
[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[3]Faculty of Science, University of Brunei Darussalam, Brunei Darussalam
[4]Department of Computer Science, Nanjing University, Nanjing, China
{Jingli.Lu, R.Wang}@massey.ac.nz; liyanagecd@yahoo.co.nz; gaoy@nju.edu.cn

ABSTRACT

Unlike conventional automatic continuous speech segmentation models that deal with each boundary time-mark individually, in this paper, we propose an interval-data-based Linear Regression Model for syllable nucleus Durations Estimation (LRM-DE), which treats syllable boundary time-marks in pairs. This characteristic of LRM-DE makes it more suitable for estimating syllable durations for English sentences, which can be used for sentence stress detection. LRM-DE combines the outcomes of multiple base automatic speech segmentation machines (ASMs) to generate final boundary time-marks that minimize the average distance of the predicted and reference boundary[1]-pairs of syllable nuclei. Experimental results show that on TIMIT dataset, LRM-DE reduces the average difference between the predicted syllable nucleus durations and their reference ones from 13.64ms (the best result of a single ASM) to 11.81ms. Also, LRM-DE improves the syllable nucleus segmentation accuracy from 81.59% to 83.98% within a tolerance of 20ms.

***Index Terms*—** Automatic speech segmentation, multiple linear regression, ensemble model

1. INTRODUCTION

Research shows that syllable nucleus duration (i.e. vowel duration in a syllable) related features are critical for English sentence stress detection [1]. A straightforward way to estimate the syllable nucleus durations is using automatic speech segmentation technology. An automatic speech segmentation machine (ASM) is a system producing a sequence of boundary time-marks, given an utterance and its phonetic transcription.

Hidden Markov Model (HMM) based forced alignment is the most commonly used automatic segmentation algorithm. However, the HMMs in forced alignment are built mainly for identifying phonemes, not for detecting the phoneme boundaries. Thus, they can capture a certain amount of information to identify what the phonemes are, but they can only provide limited knowledge about the phoneme transition [2]. Therefore, to accurately model the phoneme transition, various improvement algorithms have been developed, which can be roughly categorized into two groups: refinement methods and ensemble methods.

In the refinement methods, some tuning techniques are used to refine the raw segmentations obtained by an ASM, such as support vector machine [2], multilayer perceptron [3], statistical correction to compensate for the systematic error [4], and the context-dependent boundary model [5]. Instead of using the maximum likelihood criterion that is used in conventional forced alignment, the minimum boundary error criterion is used in [6].

In the ensemble methods, some techniques are used to post-process the segmentation results of multiple ASMs to get the final boundary time-marks. In [7], the final boundaries are obtained by averaging the outputs of multiple ASMs. In [8,9], the final boundaries are achieved by the weighted sum of the bias-corrected boundaries.

The conventional automatic segmentation algorithms try to minimize the differences between the estimated phoneme boundaries and their reference counterparts without considering the phoneme duration differences between the estimated and the reference ones. These conventional segmentation algorithms are suitable to automatically generate time-aligned phoneme annotation of speech corpora for unit selection based concatenative speech synthesis and isolated-unit training based speech recognition. However, this is problematic to obtain the syllable nucleus durations for sentence stress detection, since an acceptable discrepancy between the estimated and reference syllable nucleus boundary time-marks does not always lead to an acceptable difference between the estimated and reference syllable nucleus durations.

Two examples are illustrated in Figure 1. The horizontal line indicates the time axis of an utterance. The two solid dots are the reference boundary time-marks of a syllable nucleus, between which is reference segment t_r.

[1] In this paper, the reference boundaries are manually segmented, which are considered as the actual boundaries.

The segments in brackets and between two triangles are estimated speech segments t_{S1} and t_{S2}, respectively.

In the example shown in Figure 1 (a), the left bracket and left triangle are equidistance from the left dot, and the right bracket and right triangle are overlapped. From this example, we can see that the boundary time-marks of brackets and triangles have the same distances with their reference counterparts. However, reference duration t_r is closer to t_{S2} (duration between the two triangles) than t_{S1} (duration in the two brackets), since t_{S1} contains the length of its reference counterpart t_r (in other words t_{S1} includes t_r), which makes it always greater than t_r.

Similarly, in Figure 1 (b), t_{S1} lies inside in its reference counterpart t_r (in other words t_{S1} is included in t_r), which makes it is always less than t_r. t_{S2} is overlapped with t_r, which makes the difference between t_{S2} and t_r smaller than the difference between t_{S1} and t_r.

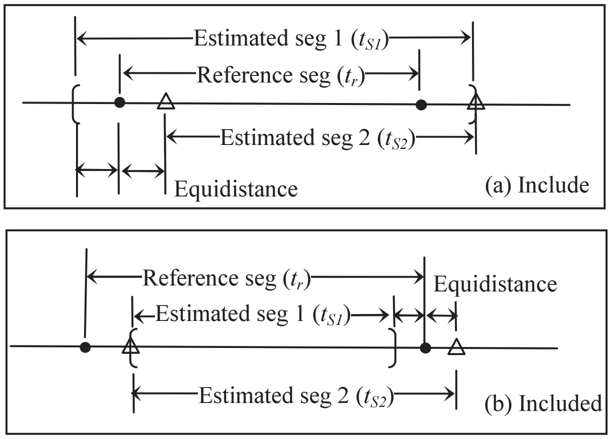
Estimated seg 1 (t_{S1})
Reference seg (t_r)
Estimated seg 2 (t_{S2})
Equidistance
(a) Include
Reference seg (t_r)
Equidistance
Estimated seg 1 (t_{S1})
Estimated seg 2 (t_{S2})
(b) Included

Figure 1. Comparison between estimated duration and its reference counterpart

A good duration estimation model should avoid the situations that the estimated durations include (or are included in) their reference counterparts. The four possible types of relationships between the estimated and the reference durations are illustrated in Figure 2, where the dots are the reference boundaries and the brackets are the estimated boundaries. In Figure 2 (a) and (b), the estimated duration includes and is included in their reference counterparts, respectively. There is 50% chance that the situations in Figure 2 (a) and (b) occur, if the automatic speech segmentation machine only tries to minimize the difference between boundary time-marks.

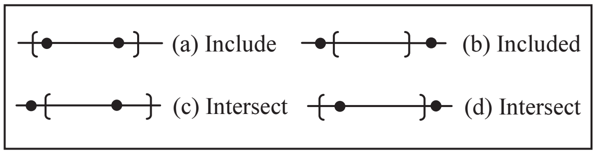

Figure 2. Relationships between the estimated and the reference syllable durations

To reduce the chance that an estimated duration includes (or is included in) its reference counterpart in automatic speech segmentation, in this paper, we propose an ensemble ASM that minimizes the discrepancy of both the estimated boundaries and durations with their reference counterparts. It is an interval-data-based Linear Regression Model for syllables nucleus Durations Estimation (LRM-DE), which combines the outcomes of multiple base ASMs. In LRM-DE, we minimize the average distance between the predicted and reference boundary-pairs of syllable nuclei.

This paper is organized as follows: Section 2 describes our proposed interval-data-based LRM-DE; the evaluation and experimental results are presented in Section 3; finally, Section 4 gives our conclusions.

2. A LINEAR REGRESSION MODEL FOR DURATIONS ESTIMATION (LRM-DE)

Phoneme durations estimation will be used to describe our LRM-DE in this section, since most ASMs are phoneme segmentation machines. The whole procedure can be carried out to estimate syllable nucleus durations.

LRM-DE is a weighted sum of segmentation results of different base ASMs, which minimizes the average distance between the estimated and reference speech segments. To reduce the possibility that an estimated duration includes (or being included in) its reference counterpart, both boundaries and duration are taken into consideration in LRM-DE. Since the durations of phonemes are a part of criterion to be optimized in LRM-DE, the phonemes boundaries have to be considered in pairs (i.e. their starting and ending points).

In conventional speech segmentation systems, the starting point of a phoneme is identical to the ending point of its previous phoneme. Then, the duration of a phoneme is the segment between its starting point and that of its subsequent one's. This duration dependency expands to a whole utterance, which makes solving the boundaries and durations optimization problem computationally infeasible. To solve this problem, in our LRM-DE, the relationship of two conjunctive phoneme segments is more flexible, which can be overlapped, connected or unconnected, as shown in Figure 3.

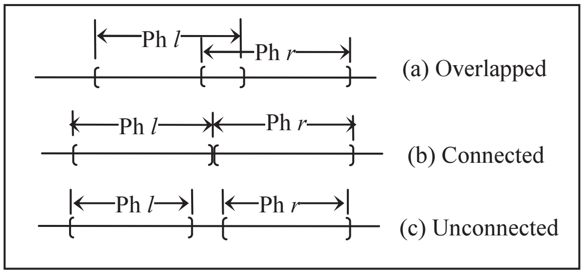

Figure 3. Boundary relationships of two conjunctive phonemes

The flexible phoneme boundaries relationships have the ability to model the relations between the durations of two conjunctive phonemes better than the simply connected phoneme boundaries relationship. It has been noticed that it is not easy to segment speech into small units consistently, even for very experienced labellers. For example, they have difficulties in marking the boundaries of vowel-to-vowel [9]. Moreover, the segmentations of different labellers may be inconsistent, and the segmentations of the same labeller in different times may be inconsistent. This is mainly because some phoneme transitions are ambiguous. For some speech segment around a boundary, it is difficult to tell which phoneme it belongs to. It may sound like both of the two phonemes or none of them. The connected boundary relationship alone cannot model the ambiguity in phoneme transitions. Thus, in our LRM-DE, it is reasonable to assume that the boundary relation of two conjunctive phonemes can be overlapped, connected or unconnected.

Given utterance *u* with its phonetic transcription and the outputs of *K* base conventional ASMs, for phoneme *p* in *u*, whose previous and subsequent phonemes are *l* and *r*, its starting and ending points generated by the *K* base ASMs are $t_j(l)$ and $t_j(p)$, where j=1,..., *K*. $t_j(l)$ is also the ending points of phoneme *l* and $t_j(p)$is also the starting point of phoneme *r* in the *K* base ASMs, since the *K* base ASMs are conventional ASMs. The time-mark interval of *p* achieved by LRM-DE is as follows,

$$[\hat{s}(p),\hat{e}(p)]=\sum_{j=1}^{K} w_{j,l-p+r}[t_j(l),t_j(p)]+[s,e]_{l-p+r} \qquad (1)$$

where $w_{j,l\text{-}p+r}$ is the weight of the j^{th} base ASM given the triphone type *l*-*p*+*r*, [s, e]$_{l\text{-}p+r}$ is the overall system error. Figure 4 gives an overview of LRM-DE.

To estimate the weights of the base ASMs and system error for each triphone model, we define the *distance* of two segments by Eq.(2), which takes both the boundary difference and duration difference into consideration:

$$\begin{aligned}&Dis([s_i,e_i],[s_j,e_j])\\&=((e_i-s_i)-(e_j-s_j))^2+(s_i-s_j)^2+(e_i-e_j)^2\end{aligned} \qquad (2)$$

The criterion to estimate the weights of the base ASMs and system error for each triphone model is to minimize the total *distance* between the predicted and reference boundary-pairs of the corresponding triphone instances. For a triphone model *l*-*p*+*r*, the objective function is:

$$\text{Min:}\sum_{p\in A} Dis([\hat{s}(p),\hat{e}(p)],[t_r(l),t_r(p)]) \qquad (3)$$

$$\hat{s}(p)=\sum_{j=1}^{K} w_{j,l-p+r}t_j(l)+s_{l-p+r}$$

$$\hat{e}(p)=\sum_{j=1}^{K} w_{j,l-p+r}t_j(p)+e_{l-p+r}$$

$$\text{s.t.} \ \sum_{j=1}^{K} w_{j,l-p+r}=1$$

where $t_r(l)$ and $t_r(p)$ are the reference ending points of phonemes *l* and *p*.

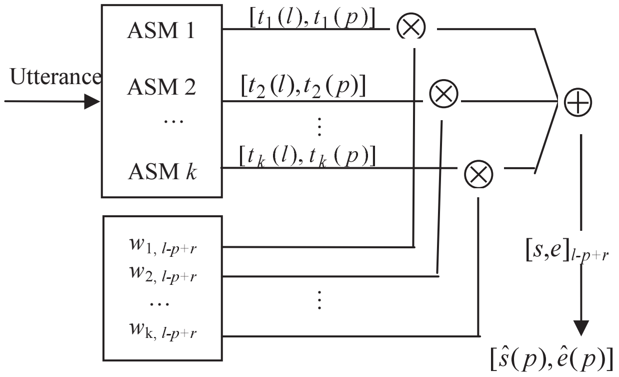

Figure 4. Overview of LRM-DE

3. EVALUATION

3.1. Experiment setup

Our experiments were conducted on TIMIT dataset, which is an acoustic-phonetic corpus of English speech. TIMIT handbook suggests the training and test sets. In our experiments, we used the training set to train HMMs and the core test set to test different base ASMs and LRM-DE. According to the suggestion in TIMIT handbook, we discard SA1 and SA2 sentences in both training set and test set, because they were uttered by both training and test speakers. Then, in our experiments, the training set and test set contain 3696 and 192 sentences respectively.

The base ASMs were trained by using the HTK toolkit [10]. The window size and frame size are set as 10ms and 5ms respectively. A 39-dimension feature vector is calculated for each frame of the speech data, which is composed of 12 MFCC with energy [11], 13 first order deviations and 13 second order deviations. The 61 phonemes in TIMIT are mapped down to 52 phonemes that are represented by left-to-right context-independent HMMs. There are 33 base ASMs that are 11 different numbers of mixtures (from 1, 2, 4, … , up to 20, step size is 2) for 3-state, 5-state and 7-state HMMs respectively.

3.2. Evaluation measurements

Seven measurements are used in the evaluation in Table 1.

MBdyErr_ph is defined as the average difference between the estimated and reference phonemes boundaries, as shown in Eq. (4).

$$MBdyErr\_ph=\frac{1}{N_P}\sum_{i\in P}(|s_{r,i}-s_{e,i}|+|e_{r,i}-e_{e,i}|)/2 \qquad (4)$$

where *P* is the phoneme set, N_P is the cardinality of *P*, $s_{r,i}$, $e_{r,i}$ ($s_{e,i}$ and $e_{e,i}$) are the reference (estimated) starting and ending points of i^{th} phone. *MBdyErr_v* is defined in a similar manner, which is the average difference between the estimated and reference syllable nuclei boundaries.

MDurErr_ph is defined as the average difference between the estimated and reference phonemes durations.

$$MDurErr\_ph = \frac{1}{N_P} \sum_{i \in P} |(e_{r,i} - s_{r,i}) - (e_{e,i} - s_{e,i})| \qquad (5)$$

Similarly, *MDurErr_v* is the average difference between the estimated and reference syllable nuclei durations.

Acc_bdy, *Acc_ph* and *Acc_v* are defined as the percentage of correctly segmented boundaries, phonemes and syllable nuclei. A phoneme or syllable nucleus is correctly segmented, if both its starting and ending points have a deviation within a tolerance with respect to its reference counterpart.

Table 1. Performances of base ASMs and LRM-DE

Measuring Quantity	Abbreviation	Best single base ASM	LRM-DE	improvement
Mean boundary errors of phonemes	*MBdyErr_ph*	9.28ms (5-st. 16-mix.)	7.58ms	1.23ms
Mean boundary errors of syllable nuclei	*MBdyErr_v*	8.81ms (5-st. 4-mix.)	8.09ms	1.19ms
Mean duration errors of phonemes	*MDurErr_ph*	12.58ms (5-st. 16-mix.)	10.70ms	1.88ms
Mean duration errors of syllable nuclei	*MDurErr_v*	13.64ms (7-st. 6-mix.)	11.81ms	1.83ms
Segmentation accuracies of boundaries	*Acc_bdy* (<20ms)	90.56% (5-st. 16-mix.)	92.70%	2.14%
Segmentation accuracies of phonemes	*Acc_ph* (<20ms)	83.35% (5-st. 16-mix.)	86.79%	3.44%
Segmentation accuracies of syllable nuclei	*Acc_v* (<20ms)	81.59% (7-st. 4-mix.)	83.98%	2.39%

3.3. Experimental results

Since the training data is limited and some triphones may meet sparse data problems, to achieve robust weights estimation in LRM-DE, we clustered triphones into different groups. The clustering method we used is the same as the method of creating tied-state triphones [10].

Table 1 shows the performances of the best single base ASMs and LRM-DE for each evaluation measurement within a tolerance of 20ms. LRM-DE increases phoneme segmentation accuracy *Acc_ph* to 86.79% from 83.35% (the best result of a single ASM). Syllable nuclei segmentation accuracy *Acc_v* is increased from 81.59% (the best result of a single ASM) to 83.98% by LRM-DE. In LRM-DE, the mean boundary errors of phonemes and syllable nuclei decrease to 7.58ms and 8.09ms, respectively; also, the mean duration errors of phonemes and syllable nuclei decrease to 10.70ms and 11.81ms.

4. CONCLUSION

Previously, weighted sum based ensemble methods have been successfully applied into automatic speech segmentation, whereas, most of them only minimize the discrepancy between the automatic segmentation boundaries and their reference counterparts, without considering the syllables (or phonemes) durations. This may cause that the estimated segments include (or are included in) their reference counterparts, which makes the estimated durations always greater (or less) than their reference counterparts.

To alleviate the problem, we proposed a linear regression based ensemble model for syllable nucleus durations estimation, LRM-DE. It combines the outcomes of multiple base ASMs and minimizes the average distance between the predicted boundary-pairs of syllable nuclei and their reference ones. Experimental results show that LRM-DE reduces the average difference between the predicted syllable nucleus durations and reference ones from 13.64ms to 11.81ms, and improves the syllable nuclei segmentation accuracy from 81.59% to 83.98% within a tolerance of 20ms.

5. ACKNOWLEDGEMENT

Gao would like to acknowledge the support for this project from the National Science Foundation of China (NSFC grant No.60775046 and No.60721002) and the National Grand Fundamental Research 973 Program of China (grant No.2009CB320700).

6. REFERENCES

[1] Xie. H, Andreae. P, Zhang, M. and Warren, P. “Detecting Stress in Spoken English Using Decision Trees and Support Vector Machines”. *Australian Computer Science Communications*, 26(7), pp. 145-150, 2004.

[2] H.-Y. Lo and H.-M. Wang, “Phonetic boundary refinement using support vector machine,” in Proc. ICASSP, 2007.

[3] K.S. Lee, “MLP-based phone boundary refining for a TTS database,” IEEE Trans. on Speech and Audio Processing, vol. 14, pp. 981–989, 2006.

[4] D.T. Toledano and A.H. Gómez, "Automatic phonetic segmentation," IEEE Trans. Speech Audio Process., vol.11, no.6, pp.617–625, Nov. 2003.

[5] L. Wang, Y. Zhao, M. Chu, J. Zhou and Z. Cao, “Refining Segmental boundaries for TTS Database using fine contextual-dependent boundary models”, in *Proc* ICASSP, pp. 641–644, 2004.

[6] J.-W. Kuo and H.-M. Wang, “Minimum boundary error training for automatic phonetic segmentation,” in Proc. Interspeech, 2006.

[7] J. Kominek and A. W. Black, “A family-of-models approach to HMM-based segmentation for unit selection speech synthesis,” in Proc. ICSLP, 2004.

[8] S. Jarifi a, D. Pastor and O. Rosec, “A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis”, Speech communication, vol. 50, pp. 67-80, 2008.

[9] S. S. Park and N. S. Kim, “On Using Multiple Models for Automatic Speech Segmentation”, IEEE Trans. On Audio Speech and Language Processing, 15, pp.2202-2212, 2007.

[10] S.Young, G. Evermann, D.Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P.Woodland, The HTK Book. Cambridge, U.K.: Cambrige Univ, 2006.

[11] T.L. New, S.W. Foo and L.C. De Silva, “Speech Emotion Recognition Using Hidden Markov Models”, Speech Communications 41(4), pp.603-623, 2003.