SYLLABLE: A SELF-CONTAINED UNIT TO MODEL PRONUNCIATION VARIATION

Raymond W. M. Ng, Keikichi Hirose

Graduate School of Information Science and Technology, The University of Tokyo, Japan.
{reimondo,hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we demonstrate the potential of incorporating syllable-level information in acoustic modeling. The unit of syllable is not rigorously defined, which leads to a problem for its use. In this study, we derive syllable structures from the sonorant-band intensity profile of speech signal. We analyze the error statistics of a phone-based context-dependent speech recognizer and find interesting error patterns. Phone errors mainly occur inside a syllable but not at syllable boundaries. Pronunciation variation can thus be regarded as the replacement of phonetic elements within the time span of a solitary syllable. We apply simple rules to model the pronunciation variation phenomenon. A lexical modeling approach modifies the bi-phone transcription in the dictionary. It leads to a significant increase of phone correctness. The results shed light on a more intuitive and direct approach to model pronunciation variation within the scope of syllables.

Index Terms— Syllable, pronunciation variation

1. INTRODUCTION

Since the 1990s, phone-based hidden Markov models (HMM) have been popular for acoustic modeling in automatic speech recognition systems. A *phone* is the smallest segmental unit which bears meaningful contrasts in a language. For instance, the two English words "did you" comprise a sequence of five phones: /d/,/i/,/d/,/y/,/uw/. By breaking down the continuous flow of speech into these short units, a concise set of acoustic units with good statistical stationarity can be derived. Statistical models like HMM are used to model acoustic properties of each unit in this set.

In casual conversational speech, phonetic units are not uttered one by one in an isolated manner as it is transcribed. Using the above example of "did you", the five phones are often uttered in the form of /jh/,/y/,/uw/ [1]. This is a typical instance of *pronunciation variation*, where neighbouring units affect one another [2].

Context-dependent phone modeling is introduced to tackle this issue [3]. A phone has its acoustic properties modeled under the condition of the preceding and subsequent phones. Nevertheless, pronunciation variation is often triggered by long-span dependencies in speech. With acoustic models on the phone level, we have to deal with high-order context-dependent models such as quinphones where sparsity becomes an issue.

It was well known that human perception to speech receives heavy influence from syllables [4]. Greenberg pointed out the importance of syllable-level information for understanding the complicated pronunciation variation patterns [5]. In automatic speech recognition there have been attempts to incorporate syllable-level information in acoustic modeling [1][6][7][8]. Despite the significant progress made by these studies, many fundamental questions about the modeling of syllables are remained unsolved. For instance, there is not a consistent and precise definition on syllables in the acoustic domain of speech [9]. The syllable units can only be extracted by a casual concatenation of phones.

In this paper, we demonstrate how conventional modeling with the phonetic units can benefit from considering syllable-level information as well. A syllable extraction method is introduced. It considers not only the spectral information but also the sonorant-band intensity profile of speech. We analyze the performance of a speech recognition system conditioned on the syllable-level information, and demonstrate pronunciation variation is the result of the interactions of phonetic units within the scope of a syllable. The paper is further supplemented by the experiments in lexical modeling, indicating how these interactions of phones within the syllable can be easily modeled. Section 2 introduces the syllable extraction algorithm. The aforementioned analysis is given in Section 4, followed by the lexical modeling experiments in Section 5.

2. CONSTRUCTING PSEUDO-SYLLABLES FROM SPEECH SIGNAL

A common syllabification approach employs a speech recognizer. The sequences of recognized phones are segmented into blocks, each of them becomes a syllable. In the course of segmentation, a list of permitted and prohibited phone clusters is specified for the determination of onset and offset of syllables [10]. Nevertheless, being the subject matter to be reviewed under the study of pronunciation variation, we believe the phonemic identity of the recognized phone should not be used as a syllabification cue.

We implement an algorithm of syllabification using the temporal envelope of speech. The algorithm is based on the assumption that a hill-shaped profile on the temporal envelope signifies the full trajectory of a syllable from onset, nucleus to offset. This approach is commonly adopted in tasks such as language identification where an exact speech recognition output is not necessary [11].

2.1. Syllabification with temporal envelope

Temporal envelope is represented by the sonorant-band intensity profile. It is obtained by performing waveform rectification and low-pass filtering on the acoustic signal in the voice band. By choosing the voice band, it is expected to exclude the nasal sounds in low frequency, and the trace in high frequency regions which reflects phone type and quality. The exact value of the pass band has been studied in different experiments [9][12]. We refer to these studies and use a pass band between 300Hz to 1000Hz in our syllabification algorithm.

To trace a syllable from the continuous speech stream, a moving time window is applied to the temporal envelope and all local peaks

This project is partially funded by National Institute of Information and Communications Technology (NICT), Japan.

are identified. For each peak, a likelihood score is computed based on its height, temporal span and other criteria. Those peaks with scores higher than a threshold are considered as the syllable nuclei.

For each detected syllable nucleus, we trace towards both sides to look for syllable onset and offset boundaries. As it is assumed a syllable coincides with the hill-shape profile in the temporal envelope, we find the local minimum points on the temporal envelop to act as a delimiter for syllables. For the convenience in subsequent implementation, the exact boundary of syllables are aligned with closest phone boundaries according to the output from the speech recognizer. Figure 1 shows four syllables detected by using the temporal envelope and phone alignment information. The temporal span of these syllables are marked by thick grey horizontal lines. Due to the algorithm implementation, it is possible to have two neighbouring syllables whose spans overlap with each other. The acoustic wave signal and phone transcriptions are also included in the figure for reference.

There is no consistent definition of syllable. With a proper notation, the suprasegmental unit derived from this algorithm will be referred to as *pseudosyllable* hereinafter.

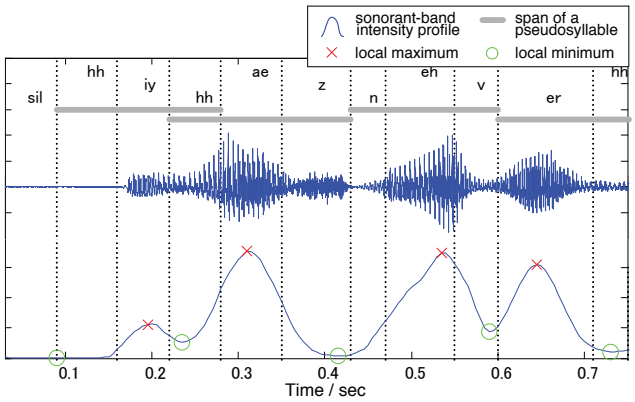

Fig. 1. *Syllabification with sonorant band intensity profile and alignment information from the speech recognizer*

2.2. Evaluation of syllabification accuracy

The accuracy of syllabification is evaluated by comparing the segmentation of syllables with some reference alignments. We used TIMIT, the read-speech corpus of English, for the comparison [13]. Out of some 4600 utterances in the training database, 204 utterances are randomly chosen. Theoretically a syllable has structure of C^nVC^m, which means an onset cluster of n consonants is followed by a vowel and then a coda with m consonants. Nevertheless, in a continuous speech stream, ambiguity arises when consonant clusters are to be divided into two syllables. We apply a rule to enforce C^nV segmentation in the generation of reference syllable alignments. C and V are derived from the phone-level forced alignment provided with the database. On top of this, aural inspection to each utterance is carried out before the exact reference syllable alignments are determined.

The insertion and deletion rates of the syllabification algorithm are 4.26% and 17.54% respectively. They reflect the percentage errors in the whole population of reference syllables. Insertion is the case where more than one pseudosyllabic nucleus is detected when the reference alignment indicates only one syllable. Deletion is the case when no pseudosyllabic nucleus is found within the reference alignment of a syllable.

3. TEST DATA AND THE SPEECH RECOGNITION SYSTEM

The speech data used is from the TIMIT database. We use the core test set with 192 utterances covering 24 speakers from 8 dialect regions [13]. Syllabification algorithm described in Section 2 is applied on the utterances. In total 1785 pseudosyllables are obtained.

The context-dependent phone recognizer is trained with the 3696 English read-speech utterances from the training set of the TIMIT database. Context-dependent triphone models which cover 39 English phones are trained. Optimal clusters of states are found by trying different clustering thresholds. Each tied state is modeled by 16 Gaussian mixtures.

With the trained triphone models, dynamic programming implemented by Viterbi algorithm using an all-phone network is applied to decode the test speech utterances. Different word insertion penalties and grammar scale factors are tried. Although the target to study are the 1785 pseudosyllables, dynamic programming is applied on the whole speech utterance to avoid adverse effects caused by abrupt speech truncation at syllable boundaries. It gives phone correctness of 61.16% and accuracy of 56.58%. This error statistic is referred to as *baseline ASR statistics* hereinafter.

4. ANALYSIS TO BASELINE ASR STATISTICS

In this section, we analyze the baseline ASR statistics conditioned on the information of pseudosyllabic boundaries. Pronunciation variation will be reviewed by comparing the segmentation mismatch and phonetic unit correctness in different parts of the pseudosyllable.

4.1. Segmentation mismatch

Segmentation mismatch measures whether a phonetic segmentation given by the recognizer output coincides with the forced alignment. From the reference transcription, 6509 phonetic boundaries are obtained with forced alignment. Some of these boundaries are also pseudosyllable boundaries as determined by the syllabification algorithm in Section 2, while others are boundaries of the phonetic units inside a pseudosyllable. We count the mismatch rate for the two kinds of phonetic boundaries (at syllable boundary and within syllable), as well as the overall mismatch rate. A 20ms tolerance is imposed when judging mismatch.

Table 1 shows the segmentation mismatch statistics. The mismatch rate is 10.7%. In other words, 89.3% of the 6509 reference phonetic boundaries can be correctly assigned by the speech recognizer. The number of phonetic boundaries within pseudosyllables and at pseudosyllabic boundaries are almost equal. However, the former has a significantly higher mismatch rate (18.2%) than the latter (3.2%).

To trace the source of segmentation mismatch, we compare the phonetic segmentation given by the recognizer output and forced alignment. While there are 6509 boundaries from forced alignment, only 6165 boundaries are detected by the speech recognizer. This suggests segmentation mismatch is caused by deletions, rather than replacements, of boundaries in the speech recognizer output.

Table 1. *Segmentation mismatch*

	Overall	at pseudo-syllabic boundary	Within pseudo-syllable
Mismatched/Total	697/6509	105/3260	592/3249
Mismatch rate	10.7%	3.2%	18.2%

4.2. Phone recognition at different parts of the pseudosyllable

To understand the recognition correctness of different phonetic units in a pseudosyllable, we draw the relevant statistics of phones conditioned on their positions in the pseudosyllable. The first and last phones in the pseudosyllable are referred to as *the phone at pseudosyllable onset* and *the phone at pseudosyllable coda* respectively. All other phones are regarded as *intra-pseudosyllabic phones*. Table 2 shows the correctness of the three types of phones in the testing database. It can be seen that phones away from syllable boundaries have significantly lower correctness than those at syllable onset or coda. One possible reason is that vowels, which locate inside the pseudosyllabic nuclei away from the boundaries, tend to suffer from severe confusion. Another reason is that a significant number of deletions are found within the pseudosyllable, which directly affects the correctness.

Table 2. *Correctness of different types of phones*

	at pseudo-syllabic onset	at pseudo-syllabic coda	Intra-pseudosyllabic
Correctness	71.5%	73.9%	52.1%

5. MODELING PRONUNCIATION VARIATION BY LEXICAL MODELING

Pronunciation variation can be modeled by replacement of phonetic elements [2]. The results given in Section 4.1 and 4.2 establish the fact that errors in a context-dependent triphone recognizer are commonly located inside a pseudosyllable. In this section, we use a lexical modeling approach to model pronunciation variation. Phone replacements are assumed to fall only inside a solitary syllable.

5.1. Alternative transcriptions

The first step of the lexical modeling is to construct a pronunciation dictionary. The dictionary maps pseudosyllable, as a structurally integral unit, to phonetic transcriptions. Given the pseudosyllable alignments, phonetic transcription can readily be generated by the concatenation of the component phones. These transcriptions are referred to as *reference transcriptions*.

To model pronunciation variation, the dictionary is expanded by some *alternative transcriptions*. *Alternative transcriptions* are found by running the speech recognizer described in Section 3 on the 3696 training utterances in the TIMIT database. Let $\boldsymbol{X}$ denote the reference transcription. If the number of times for a mis-recognition pattern (e.g. from $\boldsymbol{X}$ to $\boldsymbol{Y}$) to occur exceeds an *occurrence threshold*, $\boldsymbol{Y}$ is added into the dictionary as an *alternative transcription*. The smaller the occurrence threshold, the more alternative transcriptions are added to the dictionary.

As a preliminary attempt of lexical modeling, we confine $\boldsymbol{X}$ to be a bi-phone sequence within the pseudosyllable. To provide a brief understanding, in Table 3 we include the 41 bi-phone confusion patterns found by setting the occurrence threshold to 50. Using such list of confusion pairs, alternative transcriptions can be generated from the reference transcriptions to expand the dictionary.

Table 3. *Frequent confusion patterns from the analysis of training database (occurrence threshold=50)*

Reference transcription → Alternative transcription												
aa-r	→	er	ao-r	→	er	ih-s	→	ah-s	r-ih	→	er
ae-n	→	ae	d-b	→	b	k-t	→	k	r-ih	→	r
ae-n	→	ah-n	dh-ah	→	ah	l-ih	→	l-ah	s-ih	→	s
ae-n	→	ih-n	d-ih	→	ih	n-d	→	n	s-s	→	s
ah-l	→	l	d-ih	→	t-ih	n-t	→	n	s-t	→	s
ah-m	→	m	f-ao	→	f	n-t	→	t	t-ih	→	d-ih
ah-n	→	ah	hh-ih	→	ih	p-r	→	p	t-ih	→	ih
ah-n	→	ih-n	hh-w	→	w	r-ah	→	er	t-s	→	s
ah-n	→	n	ih-n	→	ah-n	r-ah	→	r	t-t	→	t
ah-s	→	ih-s	ih-n	→	ih				z-s	→	s
ah-v	→	ah	ih-n	→	n						

According to Table 3, 32 out of 41 confusion patterns are phone deletion from two phones to one inside a pseudosyllable. These rules are related to the deletion of phone boundaries reported in Section 4.1. Vowel deletion occurs in CV, VC and CC constructions. Particularly, the CC bi-phones such as "s-s" and "t-t" are believed to capture a cross-word pattern where the last phone in a word merges with the first phone in the subsequent word, with a high probability.

5.2. Phone recognition with updated dictionaries

The phone recognizer described in Section 3 is repeated with the use of dictionary. We construct dictionaries which incorporate alternative transcriptions under different occurrence thresholds. The phone correctness and accuracy are included in Table 4.

Table 4. *Speech recognizer performance with dictionary*

	Baseline ASR statistics	Occurrence threshold 20	5	3	2	1
Confusion pairs	N/A	207	1278	2435	4245	11817
Correctness	61.16%	66.05%	71.06%	72.29%	73.15%	73.09%
Accuracy	56.58%	60.92%	65.74%	66.98%	68.33%	69.74%

With smaller occurrence thresholds, a larger number of confusion pairs are found. When every single occurrence of mis-recognition in the training data set is considered (occurrence threshold $= 1$), phone correctness and accuracy of the testing set would be increased by more than 10% absolutely compared with the baseline ASR statistics.

5.3. An oracle test with a perfect dictionary

Finally, we conduct an oracle experiment, where hypothesis transcriptions are directly generated from confusion analysis with the test set instead of with the training set. This oracle setting essentially assumes a known pattern of confusion for every test pseudosyllable. The updated correctness and accuracy are 87.99% and 85.92% respectively.

6. DISCUSSION

6.1. A self-contained pseudosyllable to model pronunciation variation

This study suggests that pronunciation variation is a phenomenon which operates within the time span of a solitary syllable. To achieve this conclusion, a tailor-made syllabification algorithm is employed. It assumes a hill-shaped profile on the temporal envelope coincides with the pseudosyllable. Without a generally agreed definition of syllables, this assumption can only be regarded as an implementation preference which fits better with our linguistic intuitions to syllables.

Nevertheless, analysis result with the baseline ASR statistics shows this method of syllabification gives rise to a speech unit whose recognition correctness is high at the boundary and low in the middle. If we consider pronunciation variation as sequential events occurring in unknown places throughout the continuous speech stream, we will immediately find that pseudosyllable boundaries act as a perfect delimiter to these events. Pronunciation variations can be boiled down to a simple phenomenon described by phone replacements within one pseudosyllable. Thus, we suggest that pseudosyllable is a *self-contained* unit to model pronunciation variation.

6.2. Simple rules to model pronunciation variation

In this study, we take a lexical modeling approach to supplement the dictionary with alternative transcriptions. When an alternative transcription is generated, the edit distance from the corresponding reference transcription is always limited to one bi-phone (i.e. a pronunciation variation rule in Table 3 is applied only once in the generation of an alternative transcription). We have tried relaxing this constraint and allow multiple pronunciation variation rules to be applied. This increases the size of dictionary. Yet, the best correctness attained is 74.09%, as compared with 73.09% in Table 4 with the single edit-distance constraint. This indicates the pronunciation variation in the pseudosyllable is a simple operation involving the replacement of a limited number of phonetic units within the pseudosyllable.

7. SUMMARY AND FUTURE WORK

In this paper, we demonstrate the potential of incorporating syllable-level information in acoustic modeling. By considering syllable-level and phone-level information together, we find interesting phone recognition error patterns. We suggest that pronunciation variation can be described by the interaction of phonetic units within the scope of a pseudosyllable, and it can be modeled by simple rules. The results in this experiment shed light on a simple approach to model pronunciation variation within the scope of syllables.

The oracle experiment in Section 5.3 shows the capability of using alternative transcriptions when it can be generated in a perfect manner. The following question to ask is how to generate these alternative transcriptions. Long-distance dependencies across pseudosyllables may play a part here. On the other hand, an option other than expanding the pronunciation dictionary would be to refine the phonetic inventory. By doing so, we are expected to end up with a more faithful acoustic model where the statistical properties of speech signal is better reflected.

8. REFERENCES

[1] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G.R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358 –366, May 2001.

[2] M. Wester, J. M. Kessens, and H. Strik, "Pronunciation variation in ASR: which variation to model," in *Proceedings of ICSLP*, 2000, vol. 4, pp. 488–491.

[3] K.-F. Lee, "Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599 –609, April 1990.

[4] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 82–87, February 1975.

[5] S. Greenberg, "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, November 1999.

[6] H. Wu and X. Wu, "Context dependent syllable acoustic model for continuous Chinese speech recognition," in *Proceedings of Interspeech*, 2007, pp. 1713–1716.

[7] K. Thambiratnam and F. Seide, "Fragmented context-dependent syllable acoustic models," in *Proceedings of Interspeech*, 2008, pp. 2418–2421.

[8] A. Hämäläinen, L. ten Bosch, and L. Boves, "Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider," *Speech Communication*, vol. 51, no. 2, pp. 130–150, February 2009.

[9] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proceedings of ICSLP*, 1996, pp. 1261–1264.

[10] W. M. Fisher, *Syllabification software*, http://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z, June 1997.

[11] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1904–1911, August 2007.

[12] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modeling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[13] "TIMIT acoustic-phonetic continuous speech corpus," National Institute of Standards and Technology, Gaithersburg, MD, 1990.