Automatic segmentation of English words using phonotactic and syllable information

Raymond W. M. Ng, Keikichi Hirose

Graduate School of Information Science and Technology, The University of Tokyo

{reimondo,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

It is difficult to demonstrate the effectiveness of prosodic features in automatic word recognition. Recently, we applied the suprasegmental concept and proposed an extra layer of acoustic modeling with syllables. Nevertheless, there is a mismatch between the syllable and the word units and that makes subsequent steps after acoustic modeling difficult. In this study, we explore English word segmentation without a pronunciation dictionary. The algorithm is based on phonotactic and pseudosyllable information trained on a direct model with conditional random fields. An F-measure of 0.69 is attained. This result opens the possibility of automatic word recognition with the extra layer of syllable modeling.

Index Terms: automatic word recognition, word segmentation, pseudosyllable

1. Introduction

Speech prosody provides valuable information to automatic speech processing tasks. For instance, in the task of speaker and language identification, characteristics of the target speaker or language classes can be modeled with prosodic features [1][2]. The suprasegmental features are also useful by providing additional information on structures, semantics or pragmatics beyond the word level. Examples include disfluency detection, as well as sentence and topic segmentation [3].

While prosodic features are useful in generating idiosyncratic properties or high-level structural information, it is difficult to demonstrate the effectiveness of prosodic features in another important task - word recognition. In previous approaches, prosodic features were used to directly expand the feature space, or they were incorporated into backend processes like hypothesis re-ranking. Only some improvements could be obtained [3].

In our recent study, a different approach to make use of prosodic information has been proposed. Instead of expanding the feature space with prosodic features such as F0 and duration, we kept the cepstral feature space unchanged. Decoded phones were considered in blocks of syllables, which formed an extra layer on top of the phone layer. The unit of syllable is a suprasegmental unit normally used for extracting prosodic features. It was rarely considered in automatic speech processing tasks. Nevertheless, by this new hierarchy, pronunciation variation could be effectively modeled. An increase in phone recognition correctness with the TIMIT corpus from 61% to 73% was observed [4].

By introducing an extra syllable layer in automatic word recognition, we have to tackle the mismatch between syllable and word boundaries. This is illustrated in Figure 1. The first syllable "hh-iy-hh" spans across two words "He has", while the second syllable "aa-z" is a sub-word unit. Modeling pronunciation variation within a syllable was shown to give more robust acoustic model than modeling within a word [4]. Nevertheless, if we try to model pronunciation variation in a cross-word syllable simply by adding alternative pronunciations in the dictionary, we will run into troubles because one cross-word syllable may covers hundreds or even thousands of word pairs.

In this paper, we will conduct a preliminary study to investigate whether English word segmentation could be done automatically without using a pronunciation dictionary. Phonotactic and the syllable boundary information will be exploited. The problem of English word segmentation is formulated in Section 2. Conditional random fields, a direct modeling approach, will be used to tackle the word segmentation problem. Experimental details and results will be introduced in Section 3 and 4. Implications of the results, the importance of syllables and future work will be discussed in Section 5 and 6.

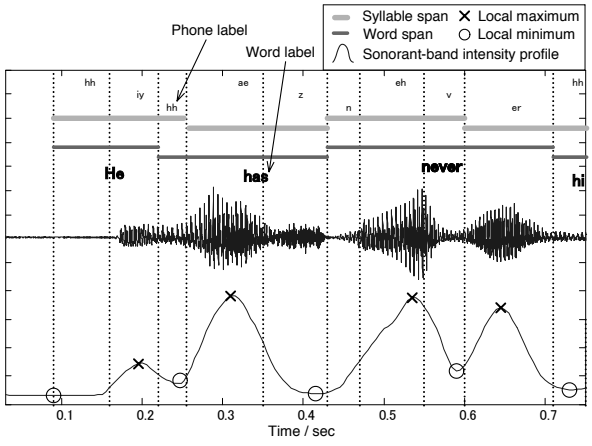

Figure 1: *The mismatch of syllable and word segmentation. Syllable segmentation is generated from the sonority-band intensity profile.*

2. Pseudosyllable to word conversion

2.1. Pseudosyllable

There is no conventional definition on syllable. Such a unit is difficult to extract. In our study, we use acoustic cues exclusively to extract a suprasegmental unit known as *pseudosyllable*. We implement an algorithm of syllabification using the temporal envelope of speech [4]. The algorithm is based on the

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

assumption that a hill-shaped profile on the temporal envelope signifies the full trajectory of the pseudosyllable from onset, nucleus to offset.

In the extraction algorithm, temporal envelope is represented by the sonorant-band intensity profile. Acoustic signal first goes through a band-pass filter, set between 300Hz and 1000Hz, to yield sonorant-band signal. Temporal envelope is obtained by waveform rectification and low-pass filtering to the sonorant-band signal. A moving window is applied to the temporal envelope and all local peaks are identified. Figure 1 illustrates the temporal envelope with the curved line at the bottom, on which the crosses mark the detected local peaks. The local peaks indicate the location of maximum energy of the vocalic portion of speech and are regarded as the nuclei of the pseudosyllables. The boundaries of pseudosyllables are determined by the valleys of the temporal envelope, and made aligned with the nearest phone boundaries according to the phone segmentation output. Details of the syllabification algorithm is explained in [4].

In this preliminary study on pseudosyllable and word, we used the forced alignments instead of the phone recognition results to find pseudosyllable boundaries. Hence, syllables and words can be interpreted as phone sequence segmentation with different criteria. For example, the first syllable "hh-iy-hh" consists of three phones while the first word "He" consists of two phones only.

2.2. Problem formulation

In conventional English word recognition approaches, a pronunciation dictionary is used to generate deterministic mappings between the decoded phone sequence and words. As explained in Section 1, this does not work for our approach with the extra layer of pseudosyllable. We refer to word segmentation for languages like Chinese, where word boundaries are not explicitly labeled in orthography [5]. Conditional random fields (CRF) are shown to be a robust approach for this task [5][6].

Under this approach, word segmentation is modeled as a labeling task. Consider two random variables $\mathbf{X}$ and $\mathbf{Y}$ which respectively range over a phone sequence and a label sequence. $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$ represents N phones. Segmentation is realized by an assignment of labels $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N)$ to each phone. The label is binary valued, having the value 1 when the phone is on the word offset and 0 otherwise.

The conditional random field is an undirected graphical model globally conditioned on the phone sequences $\mathbf{X}$ [7]. The graphical model allows arbitrary structures of the graph which represents label sequences $\mathbf{Y}$. $\mathbf{Y}$ is often assumed to be a chain, and the joint distribution over a particular label sequence $\mathbf{Y} = \mathbf{y}$ given a phone sequence $\mathbf{X} = \mathbf{x}$ is described by,

$$P_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp\left(\sum_{n=1}^{N}\sum_{k} \lambda_k f_k(y_{n-1}, y_n, \mathbf{x})\right) \quad (1)$$

Literally, Eq.(1) says that the posterior probability of $\mathbf{y}$ conditioned on $\mathbf{x}$ is calculated with a log-linear model, where the terms $f_k()$ are summed together with weight λ_k. $Z_\mathbf{x}$ is a normalization term that guarantees different label sequences $\mathbf{y}$ have their posterior probability summed to 1. $f_k()$ are known as *feature function*, it specifies a relationship between $\mathbf{x}$ and $\mathbf{y}$ and is usually binary-valued. For instance, a feature function $f_{k=k_1}()$ might be true if the phone $\mathbf{X}_n$ is a voiced bilabial plosive (i.e. /b/) and $\mathbf{Y}_n = 1$ (i.e. at word offset). Another feature $f_{k=k_2}()$ might be true if the phone $\mathbf{X}_n$ is at the pseudosyllable boundary and $\mathbf{Y}_n = 1$. By adopting different criteria, a large number of feature functions will be available to model the relationship between $\mathbf{X}$ and $\mathbf{Y}$.

In the model training process, the values of λ_k for all feature functions are trained to maximize the conditional log-likelihood of a given training set. With the trained parameters, dynamic programming finds the most likely sequence of $\mathbf{Y}$ to accomplish the word segmentation process.

3. Experiment

The data used is from the TIMIT database. For the training set, 3696 utterances which cover 4910 dictionary words are used. There are a total of 29985 word tokens. In other words, each utterance is about eight words long. For the testing set, there are 192 utterances which cover 902 dictionary words. Among the 1565 word tokens, 1146 also appear in the training set. Nevertheless, training and testing utterances are mutually exclusive. There is no training-testing utterance pair with identical word composition. Only 234 of the 1373 word bigrams in the test set appear in the training set. The testing set is considered to be an open set.

Phonotactic and pseudosyllable information are extracted for conditional random field (CRF) modeling. Phone transcriptions are directly taken from the corpus. Pseudosyllables are found by the algorithm described in Section 2.1. Feature functions are established and CRF models are trained to relate these information to word segmentation. For testing, the location of word boundaries given the phone transcriptions and pseudosyllable information is determined. In the following, we will describe how feature functions are used to model the two types of information about the phonotactics and the pseudosyllables in different orders.

3.1. Phonotactic information

Four settings of CRF models are trained with their feature functions modeling the phonotactic information in different orders.

$P_{(0)}$	Transcription of the current phone
$P_{(-1...0)}$	Transcription of the current phone, the previous phone, and their interactions
$P_{(-1...1)}$	Transcription of the current phone, the previous phone, the next phone and their interactions
$P_{(-2...1)}$	Transcription of the current phone, two previous phones, the next phone and their interactions

For the $P_{(0)}$ setting, one boolean-valued feature function is specified for every unigram phone and its word offset label y_n. Recall the example given in Section 2.2, an example feature function specifies the relationship between a voiced bilabial plosive phone (i.e. /b/) and a positive word offset label (i.e. the phone is at the word offset). For the English phone inventory with 39 phones and two output labels ($\mathbf{Y}_n = 1$ and $\mathbf{Y}_n = 0$), there will be $39 \times 2 = 76$ feature functions. The $P_{(-1...0)}$ setting uses feature functions to specify the current phone, the previous phone, as well as their interactions. Considering also the two output labels, the total number of feature functions will be $(39 + 39 + 39^2) \times 2 = 3198$ theoretically. In practice, the specified features may not occur in the training data set, or the feature functions could be pruned by the training algorithm. The actual number of feature functions will thus be smaller. $P_{(-1...1)}$ considers the previous, the current, the next

phones and their interactions. $P_{(-2...1)}$ extends to two previous phones.

3.2. Pseudosyllable information

Apart from the phonotactic information, pseudosyllable information will also be used. With the pseudosyllable boundaries detected as described in Section 2.1, two boolean variables S^{on} and S^{off} track whether the phone is the first and the last phone of a pseudosyllable respectively. Similar to the feature functions for phonotactics, pseudosyllable information with different orders are modeled.

$S_{(0)}$	S^{on} and S^{off} of the current phone
$S_{(-1...0)}$	S^{on} of the current phone, the previous phone, and their interactions, plus S^{off} of the current phone
$S_{(-1...1)}$	S^{on} of the current phone, the previous phone, and their interactions, plus S^{off} of the current phone, the next phone and their interactions
$S_{(-2...1)}$	S^{on} of the current phone, two previous phones, and their interactions, plus S^{off} of the current phone, the next phone and their interactions

4. Results

Word segmentation will be treated as a task of detecting word offset phones ($\mathbf{Y}_n = 1$). Similar to a typical tagging task, the result is evaluated with precision, recall and F-measure. Precision measures the proportion of correctly detected phones out of all detected word offset phones. Recall measures the proportion of correctly detected phones out of the true set of word offset phones with $\mathbf{Y}_n = 1$. In other words, precision measures false positive and recall measures false negative. F-measure summaries precision and recall, with the following equation,

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

We first compare CRF settings with phonotactic information in different orders. Then, the CRF models are supplemented with pseudosyllable information.

4.1. Segmentation with phonotactic information

Figure 2 illustrates the F-measure for using different CRF settings. The four thicker black bars correspond to the F-measure with four CRF settings $P_{(0)}$, $P_{(-1...0)}$, $P_{(-1...1)}$ and $P_{(-2...1)}$ where only phonotactic information is used. Table 1 included also precision and recall of these four settings.

With only the information of the current phone identity and some transition probabilities, the F-measure when using $P_{(0)}$ is 0.40. As the phone identity of the previous and the next phone are added to the model (in $P_{(-1...0)}$ and $P_{(-1...1)}$), precision and recall increase. Nevertheless, when information of one more previous phone is introduced, recall and F-measure drop. This is believed to be a data sparsity problem. In $P_{(-2...1)}$, the number of feature function is very large because up to quadrigram information is modeled. Most of the quadrigrams cannot be found in the training data and the robustness of the trained model is affected.

4.2. Addition of pseudosyllable information

Except for $P_{(-2...1)}$ which suffers from data sparsity, we introduce pseudosyllable information to the other three CRF settings $P_{(0)}$, $P_{(-1...0)}$ and $P_{(-1...1)}$. The F-measure of these augmented models are plotted in Figure 2 with thinner grey bars.

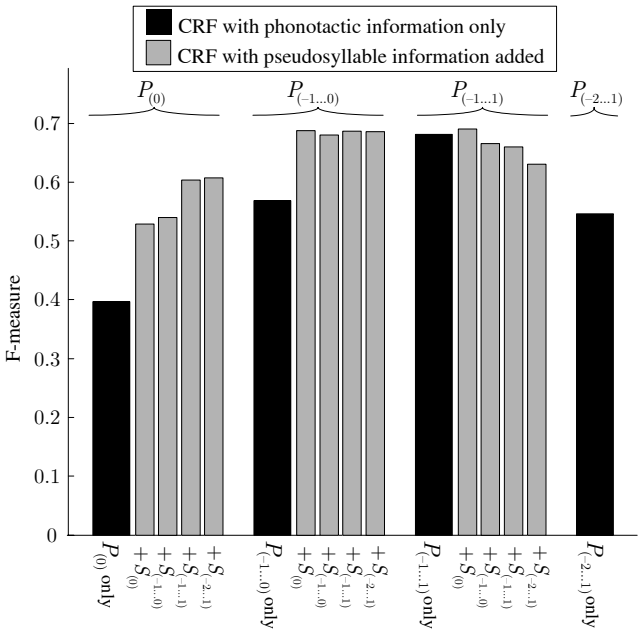

Figure 2: *F-measure with different CRF settings*

Table 1: *Performance with different CRF settings modeling phonotactic information*

	$P_{(0)}$	$P_{(-1...0)}$	$P_{(-1...1)}$	$P_{(-2...1)}$
Precision	0.53	0.59	0.73	0.79
Recall	0.32	0.55	0.64	0.42
F-measure	0.40	0.57	0.68	0.55

Table 2 lists the exact figures of F-measure for all CRF settings. The first row shows the change of F-measure when we add pseudosyllable information to the $P_{(0)}$ setting. Performance keeps improving up to the addition of $S_{(-1...1)}$, then improvement saturates when F-measure rises to a value around 0.60. From the second row of the Table, similar pattern of performance improvement is observed for the augmented models of $P_{(-1...0)}$, but performance improvement saturates at an earlier point. Moving to the third row, statistics show that adding additional pseudosyllable information to $P_{(-1...1)}$ does not seem to help significantly. The F-measure even reduces beyond $S_{(-1...0)}$.

The reason for performance to saturate (or even deteriorate) in augmented $P_{(-1...1)}$ settings as well as $P_{(-2...1)}$ setting is data sparsity. By looking into the model size of different settings, it is discovered that when the number of feature functions reaches 5000, F-measure attains its maximum at around 0.68 to 0.69. When the number of feature functions rises above 30000 (which is the case for $P_{(-1...1)}$ augmented with $S_{(-1...1)}$ and $S_{(-2...1)}$, as well as $P_{(-2...1)}$), F-measure drops. The rate of recall drops significantly to below 0.6. However, precision increases with the decreased number of word offsets detected by the models.

To summarize the result across all different CRF settings, the $P_{(-1...1)} + S_{(0)}$ setting achieves the best result in terms of F-measure (0.6899). It has a precision of 0.82 and the recall rate

is 0.60. The second best setting is $P_{(-1...0)} + S_{(0)}$, giving an F-measure of 0.6874. The precision and recall is more balanced, at a value of 0.68 and 0.70 respectively.

Table 2: *F-measure with different CRF setting with phonotactic and pseudosyllable information*

Phonotactic information	Additional pseudosyllable information				
	nil	$S_{(0)}$	$S_{(-1...0)}$	$S_{(-1...1)}$	$S_{(-2...1)}$
$P_{(0)}$	0.40	0.53	0.54	0.60	0.61
$P_{(-1...0)}$	0.57	0.69	0.68	0.69	0.69
$P_{(-1...1)}$	0.68	0.69	0.67	0.66	0.63

5. Discussion

5.1. Evaluation to results

The task of English word segmentation conducted in this study is not a typical one. It is difficult to find related studies with which we can compare our results. There is one relevant study conducted by Harrington et al. [8]. In a comparable task of word segmentation with 145 sentences, the hit and false alarm rate of word boundary detection were reported. Following Eq.(2), the equivalent F-measure of the best performing data set was 0.56. This is significantly lower than the best F-measure we report in Table 2, which is 0.69.

The performance difference between Harrington's experiment and ours is mainly due to the different approaches adopted. Harrington tried to incorporate phonotactic knowledge by looking for phone sequences which span across words yet never happen word internally. No probabilistic model was involved. The feature functions we employ in CRF actually serve the same purpose, but the CRF framework guarantees the robustness of the trained parameters.

In our experiment, both phonotactic and pseudosyllable information are used for the word segmentation task. For CRF settings which model lower-order phonotactic information (i.e. $P_{(0)}$, $P_{(-1...0)}$), F-measure increases with the addition of pseudosyllable information. Data sparsity in higher-order phonotactic models lead to performance saturation. Assuming this can be solved (e.g. by introducing more data), it is unknown whether pseudosyllable information can further boost the segmentation performance. Further experiments are necessary before a conclusion can be reached.

5.2. Importance of syllable in automatic word recognition

In automatic word recognition, phoneme or sub-syllabic segmental units are normally used as the basic unit for acoustic modeling. Syllables are rarely considered. Nevertheless, syllables are by no means a trivial unit for human in the process of speech understanding.

Word segmentation is a crucial step for human to understand the long sequences of phonetic elements that come in speech. However, signs of word boundary are not found directly in speech. To assist the perception of words, syllable information is heavily used. In a perceptual study, English speakers were found to assume strong syllable as a signature to the start of a word [9]. Actually, simple monosyllabic structures can conclude 75% of the words in the Switchboard corpus [10]. In [11], the perceptual grounds on using syllables for human speech perception were explained. It was argued that speech understanding did not require a detailed spectral portraiture of the signal.

With the above evidence from human perceptual studies, it is expected that syllable modeling should be explored for automatic speech recognition as well. However, there are a lot of problems before one can make use of syllables in automatic speech recognition with syllables. For instance, given the huge inventory, it is impossible to model syllables as we do to phones. Ganapathiraju showed that the use of a syllable/tri-phone hybrid system only slightly decreased the word error rate [10]. Thanks to our previous research in prosody, we applied the suprasegmental concept with syllables and defined a new hierarchy in modeling [4]. This approach could evade the aforementioned problem. Target units to model are still phones, just that an extra layer of syllables is proposed in the hierarchy. The result of word segmentation reported in this paper opens the possibility of automatic speech recognition using this new hierarchy.

6. Future work

The word segmentation algorithm faces a sparsity problem when the number of feature functions is too big. Some efficient methods to prune feature functions may help. On the other hand, in this paper we use transcribed data for the word segmentation problem. In practical applications, the algorithm should work with the decoded phone sequences after pronunciation modeling. That means word segmentation algorithm will have to deal with unexpected, erroneous phone sequences. More future work is needed to address these problems.

7. References

[1] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, Jul. 2005.

[2] R. W. M. Ng, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Prosodic attribute model for spoken language identification," in *Proceedings of ICASSP*, 2010, pp. 5022–5025.

[3] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson et al., Ed., pp. 105–114. 2004.

[4] R. W. M. Ng and K. Hirose, "Syllable: A self-contained unit to model pronunciation variation," to be appeared in *ICASSP 2012*.

[5] F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," in *COLING*, 2004, pp. 562–568.

[6] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proceedings of ACL*, 2005, pp. 451–458.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random field: Probabilistic models for segmenting and labeling sequence data," in *Proc. Intl. Conf. Machine Learning*, 2001, pp. 282–289.

[8] J. Harrington, G. Watson, and M. Coooper, "Word boundary detection in broad class and phoneme strings," *Computer Speech and Language*, vol. 3, pp. 367–382, 1989.

[9] A. Cutler, "Prosody and the word boundary problem," in *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, J. L. Morgan and K. Demuth, Eds., pp. 87–99. 1996.

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

[11] Steven Greenberg, "Understanding speech understanding: Towards a unified theory of speech perception," in *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception*, 1996, pp. 1–8.