Coping Imbalanced Prosodic Unit Boundary Detection with Linguistically-motivated Prosodic Features

Yi-Fen Liu [1], *Shu-Chuan Tseng* [2], *J.-S. Roger Jang*[3], *C.-H. Alvin Chen*[4]

[1] Institute of Information Systems and Applications, National Tsing Hua University, Taiwan
[2]Institute of Linguistics, Academia Sinica, Taiwan
[3] Department of Computer Science, National Tsing Hua University, Taiwan
[4]Graduate Institute of Linguistics, National Taiwan University, Taiwan

yifenliu@gmail.com, tsengsc@gate.sinica.edu.tw, jang@cs.nthu.edu.tw, alvinworks@gmail.com

Abstract

Continuous speech input for ASR processing is usually pre-segmented into speech stretches by pauses. In this paper, we propose that smaller, prosodically defined units can be identified by tackling the problem on imbalanced prosodic unit boundary detection using five machine learning techniques. A parsimonious set of linguistically motivated prosodic features has been proven to be useful to characterize prosodic boundary information. Furthermore, BMPM is prone to have true positive rate on the minority class, i.e. the defined prosodic units. As a whole, the decision tree classifier, C4.5, reaches a more stable performance than the other algorithms.

Index Terms: prosodic unit, machine learning, biased minimax probability machine

1. Introduction

Before processing spontaneous speech, a certain level of pre-segmentation of speech signal is required. While syntactic units such as phrases, clauses or sentences are operationally well-defined units for text analyses, they do not work quite well for spontaneous speech which involves complex prosodic patterning and most importantly, they contain ill-formed sentences. Thus, in the studies of conversational speech, it has been proposed to segment conversational data into intonation units (IU), which represent a piece of meaning concept under a single, coherent intonation contour. More concretely, studies of English and Mandarin intonation unit [1] [2] show that IU is frequently followed by a pause or marked by a lengthening of the final syllable, an upward shift in overall pitch level at the beginning of IU, or a perceived change on speech rhythm.

As shown in [3], IUs overwhelmingly match clauses or semi-clauses in English, Japanese, and Mandarin. Moreover, many tasks on spoken language processing focus on how to utilize prosody to retrieve the "hidden" structural information, such as sentence-like boundary detection [4], and disfluency interruption point detection [5] [6]. It is concluded that the prosodic structure of a spoken discourse to a certain degree corresponds to its syntactic structure. Furthermore, prosodic units are marked not only by pauses, but also by other prosodic means. Prosodically defined units should be more useful for decoding the meaning of natural speech than pause-defined units. We adopted the concept and the main identification criteria of IU. But we rigidly defined it as a perceptually coherent prosodic unit which can function as an immediate unit for processing Mandarin spontaneous speech. To emphasize the role of prosody, not limiting to intonation solely, we adopted a more neutral term, prosodic units (PU) [7].

In lack of a sophisticated recognizer for Mandarin spontaneous speech, we utilized a corpus which contains manually labeled boundaries of syllables and PUs. With these PU-labeled syllables, we developed a set of prosodic features manifesting the characteristics of PU. Then we built up our PU segmentation model by means of different machine learning methods given the speech signal and labeled syllable information. That is, for each syllable boundary, the possible classes a classifier constructed with prosodic features attempts to identify are (1) PU or (2) not-PU. The proposed PU detection model utilizes prosody-related features only. We intend to integrate language models at a later stage to improve the performance of our automatic speech recognition system.

Similar to the tasks on sentence-like unit detection and disfluency interruption point detection, the main problem of such binary classification tasks is the imbalanced proportion of those two classes, namely PU and not-PU. Huang et al. [8] pointed out that traditional approaches dealing with imbalanced datasets, including sampling approaches, moving decision thresholds and adjusting the cost (weight) toward bias on important minority class, are in lack of systematic treatment on both classes. In this paper, we adopted the Biased Minimax Probability Machine proposed by Huang et al. [8] in consideration of systematic foundation and three standard classifiers: the Naïve Bayesian classifier, the *k*-Nearest Neighbor (*k*-NN) classifier, the decision tree classifier C4.5 as well as the state-of-the-art classifier, Support Vector Machine.

The organization of this paper is as follows. Section 2 defines the task on prosodic unit boundary detection and describes the experiment set up for the task, with a detailed illustration of our prosodic feature set. Evaluations of different learning methods are summarized in section 3. Section 4 shows our experiment results and findings. Section 5 concludes the paper.

2. Prosodic unit boundary detection task

2.1. Task definition

As reported in previous studies on prosodically defined units, the mean length of words of English intonation unit (EIU) and Mandarin intonation unit (MIU) are 4.8 and 3.5, respectively [1] [2]. Similarly, the mean of PU length in words labeled in our training data is 4.4. As summarized in Table 1, these units are comparable across different languages and technical definitions.

Table 1. *Mean length of prosodically defined units.*

Unit	# of word
EIU	4.8
MIU	3.5
PU	4.4

Quite often, a piece of spontaneous speech stretch segmented by pauses contains several prosodic units which indicate important information of discourse structure. In other words, it also suggests that a pause-based segmentation of spontaneous speech may not be sufficient for the task of decoding the information structure of the associated speech content. For example, the content of the pause-determined speech stretch in Figure 1 contains four utterances, which correspond to four prosodic units.

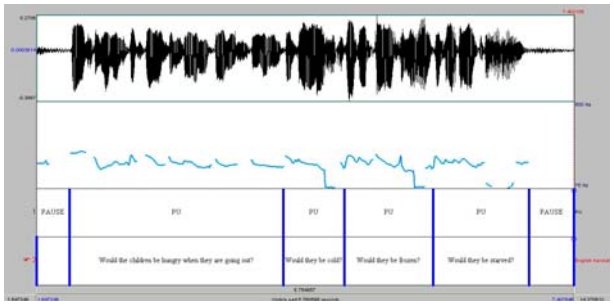

Figure 1: *A pause-determined speech stretch with waveform, pitch contour, PU labels and translations.*

Table 2 summarizes the mean length of PU and the pause-segmented unit in syllables in our dataset. PUs are apparently shorter than pause-segmented units by four syllables on average. Furthermore, PU potentially serves as a prosodic reflection of syntactic units such as sentences, clauses or semi-clauses [3]. Motivated from that, we set up this experiment for automatically segmenting long speech stretches into PU not only to enhance the recognition output but to build a more likely mapping to grammatically defined units.

Table 2. *Mean length of PU and inter-pause unit.*

Unit	# of syllables
Prosodically defined units (PU)	6.5
Pause-determined units	10.5

In experiment, we defined the problem of prosodic unit boundary detection as a binary classification task, in which each syllable boundary is viewed as a candidate for a prosodic boundary. To achieve this objective, we implemented prosodic features reflecting important linguistic characterization of PU boundaries into classifiers via different machine learning algorithms.

2.2. Experiment setup

2.2.1. Data

702 speakers turns extracted from eight transcribed conversations of the Mandarin Conversational Dialogue Corpus (MCDC) [9] were manually labeled with both syllable and PU boundaries. Among these 16 speakers, nine are female and seven are male. Table 3 summarizes the data used in this experiment, including the numbers of syllables and the percentage of PUs in the training and testing sets.

Table 3. *Data summary.*

	Train	Test
Number of PU	6450	1593
Number of not-PUs	37316	9348
PU ratio (%)	14.74	14.56

2.2.2. Prosodic features

With the labeled syllable information, we obtained a vector of prosodic features for each syllable boundary. We developed a set of 15 features, taking into account pause-related, duration-related and pitch-related characteristics of PU, mainly motivated by the feature set defined for interruption point detection in [6]. Features 01-02 directly indicate pausing. If there is a pausing, we further calculate how salient it is in comparison with the duration of the previous syllable (feature 08). For duration-related features (features 03-06), we focus on cues which manifest changes of speech rate by integrating the pause and syllable duration fluctuations. The boundary-final lengthening and boundary-initial shortening are captured in features 07 and 09. As for pitch-related features (10-15), changes of pitch values across boundaries are measured and a sequence of variants measuring pitch difference across different rhythm structure of speech is implemented. They are designed to capture an upward shifting event in overall pitch level since that the pitch jump have been identified as a useful cue for IU. Table 4 shows the complete prosodic features we used for prosodic unit boundary detection. Pitch range is defined as the difference of the maximum and minimum pitch values between the nearest pauses before and after the current syllable.

Table 4. *Prosodic features.*

No.	Description
01	Pause or not
02	Normalized pause duration based on speaker
03-06	Ratio for the average syllable length of the next N following syllables to current syllable length (N is equal to 1, 2, 3 or to N syllables before the next pause)
07	Ratio of current syllable length to average length of current syllable and the next 2 syllables
08	Ratio of pause duration to current syllable length (in msec.)
09	Ratio of current syllable (S_i) length to the average syllable length of S_{i-1} and S_{i-2}
10	Ratio of raw F0 difference to pitch range
11	Ratio of linearized F0 difference to pitch range
12	Ratio of mean F0 difference on 3 pitch values at the beginning of current 1 syllable and the next syllable to pitch range
13	Ratio of mean F0 difference on 3 pitch values at the beginning of current 2 syllables and the next syllable to pitch range
14	Ratio of mean F0 difference on 3 pitch values at the beginning of current 3 syllables and the next syllable to pitch range
15	Ratio of F0 difference of the pitch value after previous pause and the pitch value after current syllable boundary to pitch range

2.2.3. Evaluation measures

The performance of different learning techniques of this experiment is evaluated by the following two criteria.

- Maximum Sum (MS): Maximum sum manifests the sum accuracy on the positive class and the negative class.
- F-measure: The F-measure is defined as

$$F-measure=\frac{2 * \text{recall} * \text{precision}}{1 * \text{recall} \quad \text{precision}} \tag{1}$$

The precision rate is defined as TP/TP+FP, and the recall rate as TP/TP+FN (TP and FP are the number of true positives and false positives, also known as false alarms). FN denotes the number of true negatives (false misses). In this measure, the scalar term we used is a standard one to avoid imposing different cost on two different types of error.

3. Learning techniques

Here we review three methods that have been shown with a promising achievement on a balance between true positive and false positive rates: Naïve Bayesian classifier, nearest neighbor method and the decision tree classifier, C4.5 [10]. The concept and representation behind those three supervised learning methods vary. With the associated learning mechanism, they target at different degree of inductive biases on data. Generally, they all favor the majority class. Nonetheless, the Naïve Bayesian classifier has shown a better trade-off between true positive and false positive rates.

- Naïve Bayesian classifier: This method classifies a new instance according to its computed posterior probability of two classes in which it is assumed that individual features are statistically independent and then predicts the class with the highest posterior probability.
- *K*-nearest neighbor classifier: This approach classifies a new coming data by finding the *K*-nearest neighboring instances with distance measure and then assigns it with the same label associated with the majority of the *K*-nearest neighbors.
- Decision tree classifier, C4.5: In training, a decision tree recursively selects the attribute with the best value to separate the training samples into different classes. To classify a new instance, it starts from the root of the tree and follows the route determined with the pre-trained attribute value and then reaches a leaf node. At the end, the procedure returns the class label of the leaf node as the predicted class.

In addition to the above standard learning techniques, we also tested the Biased Minimax Probability Machine (BMPM) proposed by Huang et al. [8], trying to handle the learning from imbalanced data. Furthermore, we compared these models with a state-of-the-art classifier, the Support Vector Machine. SVM aims to seek a hyperplane to separate two classes of data with the maximum margin. However, the goal of BMPM is to construct the classification hyperplane by directly controlling the lower bound of real accuracy of the future data. As shown in the following formula, BMPM tries to seek a hyperplane that maximizes the accuracy (α) of the important class **x** while keeping the accuracy (β) of less important class **y** acceptable:

$$\max_{\alpha,\beta,b,\mathbf{a}\neq 0} \alpha \quad s.t. \quad \inf_{x\in\{x,\sum_x\}} \Pr(\mathbf{a}^T\mathbf{x}\geq b)\geq\alpha \tag{2}$$

$$\inf_{y\in\{y,\sum_y\}} \Pr(\mathbf{a}^T\mathbf{y}\geq b)\geq\beta \tag{3}$$

$$\beta\geq\beta_0 \tag{4}$$

For implementing aforementioned learning algorithms for PU boundary detection, we used WEKA package [11] for the three standard learning techniques and the released algorithm implemented in MATLAB code for BMPM [12]. Then we employed a support vector machine (SVM) with a radial basis kernel using LIBSVM [13].

4. Experiment results and discussion

In our PU-segmented conversational data, PUs are often followed by pauses. As pause often serves as an effective means for speakers to indicate endings of his/her speech and pause can be easily detected from acoustic information, it has come to be a good indicator of segmentation for most NLP engineering tasks. Therefore, we calculated the proportion of those PU syllable boundaries followed by a pause in the testing dataset and take the ratio as our baseline in comparison to those five learning techniques. This baseline represents the percentage of all the prosodic units one can retrieve by pauses only.

Table 5. *PU detection results in MS and F-measure*.

Learning methods	Evaluation methods			
	Maximum sum (%)			F-measure
	TP	TN	(TP+TN)/2	
Baseline	58.9	-	-	-
Naïve Bayesian	61.7	99.8	80.8	0.76
K-NN (5)	72.0	98.7	**85.4**	0.80
C4.5	71.6	99.1	**85.4**	0.81
BMPM	**74.5**	92.3	83.4	0.68
SVM	62.7	99.3	81.0	0.75

As shown in Table 5, all five learning methods are successful in identifying prosodic units within pauses, compared to the baseline TP. In other words, they all show significant improvement in segmenting conversational discourse into smaller units than pauses. We noticed that most of the incorrectly classified errors are PU boundaries which are predicated as not-PU. The PU boundaries were labeled based on the perceived cues. Possibly, the fine nature of the human perceptual judgment cannot be captured by our current feature set of acoustic cues. However, with only 15 linguistically motivated prosodic features, our model is parsimonious in nature as compared to other engineering-based detectors. Although the BMPM classifier performed the worst in terms of F-measure, it outputs the highest TP and recall rates on PU, which is the more important class.

Huang et al. [8] has argued that BMPM out-performances the Bayesian classifier, the *K*-Nearest Neighbor (*K*-NN) classifier and the decision tree classifier, C4.5. These three learning methods are all modified by changing the cost matrices or other methods introduced in [10]. For our

experiment, we used the standard classifiers without modifying any cost on the PU and not-PU classes. Our results show that both the *K*-NN classifier and C4.5 raised the accuracy in Maximum sum. In terms of the overall system performance, the C4.5 classifier performs the best, because it achieves the highest F-measure and Maximum sum.

Figure 2 illustrates part of the result obtained from the decision tree trained by our data. Examining the features selected from the decision tree learning algorithm, we found that the most salient features correspond to three main groups of prosodic cues: pausing, speech rate change and pitch reset, which are also often used for perceptually identifying prosodic units in linguistic analysis. As listed in Table 4, feature 01 (*Pause*) indicates the presence of pause. Features 06 (*RatioDurNxtNSyl2Cur*) and 07 (*RatioDurCurSyl2CurAndNxt2*) reflect the boundary effects of initial shortening and final lengthening of PU, in which local change of speech rate is modeled. In the fifth level of the illustrated decision tree, feature 11 (*RatioLinearF0DiffatCur1SylandNextSyl2F0Range*) is designed to capture an abrupt pitch change locally. The main linguistic characteristics of prosodic units are successfully quantified and learned by the feature set we proposed.

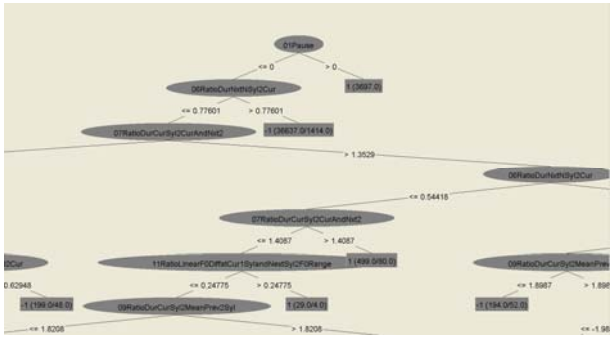

Figure 2: An example of a decision tree for all 16 speakers.

5. Conclusion and future work

This study originated from the motivation of further segmenting pause-determined speech stretches into smaller prosodic units. We used PU as a unit of coherent prosodic phrasing relying on salient prosodic cues. Aside from pauses, a set of linguistically motivated prosodic features considering pitch change and speech rhythm is developed to capture various characteristics of PUs in Taiwan Mandarin. Based on this parsimonious feature set, we built a PU detector using different machine learning techniques and evaluated their performance. The performances of all learning algorithms exceed our baseline, implying that PU is a realistic unit in spontaneous speech which at the same time is important for decoding discourse structure. Specifically, the decision tree classifier exhibits stable performance in different evaluations. So far, we only consider a single classifier as the model for PU detection. In the future, we aim to build a model combining multiple weak learning algorithms to reduce the number of two types of error. Furthermore, given our promising results on PU automatic detection, we would like to enhance the recognition output in the second pass speech recognition with the hypothesized PU boundaries as Lin et al. [14] did in their work. They show that the character accuracy of the conversational corpus is improved from 44.3% to 46.3% with the extra information of interruption point identification. Therefore, it is expected that these hypothesized prosodic unit boundaries will provide more word-level information such as cue phrases, which can be incorporated into the language model, and accordingly improve the automatic recognition system for spontaneous Mandarin speech.

6. Acknowledgements

This work was supported by the Computational Linguistics and Chinese Language Processing Program, Academia Sinica, Taiwan.

7. References

[1] Du Bois, J. W., Schuetze-Coburn, S., Cumming, S. and Paolino, D., "Outline of Discourse Transcription", in Talking Data: Transcription and Coding in Discourse Research, 45-89, 1993.

[2] Tao, H. Y., "Units in Mandarin Conversation: Prosody, Discourse and Grammar", Amsterdam: John Benjamins, 1996.

[3] Iwasaki, S. and Tao, H. Y., "A Comparative Study of the Structure of the Intonation Unit in English, Japanese, and Mandarin Chinese", in Annual Meeting of the Linguistic Society of America, Los Angeles, CA., 1993.

[4] Liu, Y., Chawla, N. V., Harper, M. P., Shriberg, E. and Stolcke, A., "A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech", Computer Speech and Language, 20:468-494, 2006.

[5] Liu, Y., Shriberg, E., Stolcke, A. and Harper, M., "Using Machine Learning to Cope with Imbalanced Classes in Natural Speech: Evidence from Sentence Boundary and Disfluency Detection", in Proceedings of ICSLP, 1525-1528, 2004.

[6] Lin, C. K., Tseng, S. C. and Lee, L. S, "Important and New Features with Analysis for Disfluency Interruption point (IP) Detection in Spontaneous Mandarin Speech", in Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech, 117-121, Aix-en-Provence, 2005.

[7] Liu, Y. F. and Tseng, S. C., "Linguistic Patterns Detected through a Prosodic Segmentation in Spontaneous Taiwan Mandarin Speech", Linguistic Patterns in Spontaneous Speech, 147-166, Institute of Linguistics, Academia Sinica, 2009.

[8] Huang, K. Z., Yang, H. Q., King, I. and Lyu, M. R., "Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine", in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 11:558-563, 2004.

[9] Tseng, S. C., "Processing Spoken Mandarin Corpora", Traitement automatique des langues, Special Issue: Spoken Corpus Processing, 45(2):89-108.

[10] Maloof, M. A., Langley, P., Binford, T. O., Nevatia, R. and Sage, S., "Improved Rooftop Detection in Aerial Images with Machine Learning", Machine Learning, 35:157-191, 2003.

[11] Hall, M., Frank, E., Homes, G., Pfahringer, B. and Witten I. H., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, 11(1), Online: http://www.cs.walkato.ca.nz/ml/weka/, 2009.

[12] Yang, H. Q., Huang, K. Z., King, I. and Lyu, M. R., "Matlab Toolbox for Biased Minimax Probability Machine" (BMPM-1.0), Online: http://www.cse.cuhk.edu.hk/~miplab/mempm_toolbox/index.htm, 2004.

[13] Chang, C. C. and Lin, C. J., "LIVSVM: A Library for Support Vector Machines", Online: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[14] Lin, C. K. and Lee, L. S., "Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech", IEEE Trans. on Audio, Speech and Language Processing, 17(7):1263-1278, September 2009.