Automatic Segmental and Prosodic Labeling of Mandarin Speech Database

Fu-chiang Chou[1], Chiu-yu Tseng[2] and Lin-shan Lee[1]

Dept. of Electrical Engineering, National Taiwan University[1]
Institute of Linguistics, Preparatory Office, Academia Sinica[2]
Taipei, Taiwan, Republic of China
email-addr: moza@speech.ihp.sinica.edu.tw

ABSTRACT

In this paper we describe the techniques and methodology developed for automatic labeling of segmental and prosodic information for the Mandarin speech database. There are two major procedures. First, the text is converted into the phonetic network of possible pronunciations, and this network is aligned with the speech data by recognition processes. Secondly, many acoustic prosodic features are derived and the break indices are labeled with these features by decision trees. For the segmental labeling, 96.5% of automatically determined segment boundaries are accurate within a range of 20 ms. For the prosodic labeling, 84.9% of the automatic labeled break indices are the same with the manual labeled one.

1. INTRODUCTION

It is now widely recognized that progress in speech technology is dependent to a considerable extent on the quality of the speech database that are available. The coverage of vocabulary, speakers, recording conditions, etc., is crucial to the success of the system. However, all these qualities can be undermined if the reliability of their labeling is suspect. The purpose of this research is to develop a set of models and tools for the automatic labeling of speech corpora. The corpora are used for training a speech synthesis system, as well as providing source units for concatenative synthesis. Stochastic models are trained to predict timing and pitch contours from the corpora. If the corpora are not accurately labeled, then the prediction of the models and the synthesis quality will degrade considerably. This labeling must closely reflect the speech as it was actually produced, and should be based on acoustic rather than on perceptual features. Thus automatic labeling may be essential for this purpose. Two kinds of automatic labeling are studied in this paper, the segmental labeling and the prosodic labeling.

Most of the automatic segmental labeling systems use the HMM-based recognizer to do a forced alignment between the speech data and the phonetic transcription[1][2]. The two major problems are:

1. have no phonetic transcription or the transcription is incorrect.
2. have no initial segmentation for the training of HMM models

For the first problem we have a program to generate the possible pronunciation networks from the accompanying orthographic transcription, and use the HMM to recognize the real phonetic realizations. For the second problem, we use a SI HMM to do the initial segmentation, and the boundaries are corrected with a set of boundary correction rules. These can make a more reliable result than the traditional procedures.

Not like the segmental labeling, there is less literature for the prosodic labeling, especially for the automatic prosodic labeling. Moreover, most of the publications on this topic are focused on English which in very different from Mandarin in nature[3][4]. In this paper, we use the decision tree to classify the boundary. This is similar to the method described in[3] but we use a hierarchical multiple pass procedure in stead of the Viterbi process. The feature set is also with many differences. The major difficulty encountered is the coupled interaction between the lexical tonal system and the prosodic system. We try to de-couple these features with some normalization procedures. Although the results cannot be directly compare with the previous study for English, this paper is a good start point for the research of automatic prosodic labeling for Mandarin.

The experimental corpus is described in section 2. The description of the segmental labeling and the prosodic labeling is in section 3 and 4. The last section is the discussions.

2. EXPERIMENTAL CORPUS

2.1 Corpus Design

In Mandarin, syllable is a very important unit. Each character is pronounced as a syllable. There are only about 400 syllables if the tonal difference is neglected. An INITIAL/FINAL format can describe the composition of a Mandarin syllable. INITIAL is the initial consonant and FINAL is the vowel (or diphthong) part with an optional medial or a nasal ending. In theory, there are about 2000 FINAL-INITIAL and FINAL-FINAL (No INITIAL in the latter syllable) patterns in the disyllabic junctures. The speech corpus is designed to cover most of these combinations[5]. Moreover, the corpus is organized as 599 short paragraphs so as to cover many prosodic variations in reading. Six professional speakers read the corpus at a normal speaking rate in a sound proof room. If there are hesitations or mistakes, the speaker will be asked to read the sentence again until each character is correctly pronounced. This can reduce the errors for further segmentation and labeling.

2.2 Corpus Labeling

The corpus is first segmental labeled by a trained transcriber. The smallest units are the INITIAL and FINAL. In the later on prosodic labeling, phrasing and emphasis are the prime determinants. The former refers to the groupings of words in an

utterance, and the latter describes the greater perceived strength or focus of certain units. The phrasing is represented by break indices. The break index is labeled for each syllabic boundary instead of word boundary because the lack of common agreement on word boundaries. We defined six possible break related boundaries corresponding to six break indices; our scale therefore 0 to 5. The defined break indices correspond to the following boundaries: reduced syllabic boundary (0), normal syllabic boundary (1), minor-phrase boundary (2), major-phrase boundary (3), breath group boundary (4), and prosodic group boundary (5). The speech segments between the break indices than form a set of prosodic units accordingly, namely, minor prosodic phrase, major prosodic phrase, breath group and prosodic group[6].

The other labeled parameter is the level of emphasis in an utterance. It is subject to the transcriber to decide the emphasis level of whatever prosodic units when they fell the emphasis. The level is from 0 to 3 corresponding to the following levels: reduced (0), normal(1), moderate (2) and strong (3). In the initial study, the automatic labeling of emphasis is not described in this paper.

3. AUTOMATIC SEGMENTAL LABELING

3.1 Training and Alignment

For the segmental labeling, the orthographic transcription of the speech data is used as the input. The possible pronunciations for the words in each sentence are derived from a text analysis module, including establishing a set of possible pronunciations for each word, and transcribing the results into the HTK's net file format. These net files include the homographs and the pronunciation variations, for examples: (重 chong2,layers and zhong4,heavy), (風 feng1 or fong1). A Viterbi process is then performed to recognize and align the units in the speech data based on the net. The units of the HMMs are context independent INITIALs and FINALs. The INITIAL has 3 states and the FINAL has 5 states. The feature vectors include 12 dimensions of MFCC, 1 dimension of RMS power and their differential values. The frame rate is set to 5ms to increase the precision of the segmentation. The traditional training procedure for HTK is illustrated in Figure 1. The alignment results showed that the HMMs trained with this procedure have some bias for the position of boundaries. The INITIALs are too long in most of the cases. The trained model may be fine for the recognition, but the alignment is not satisfactory and required further manual adjustments. During the manual correction of the alignment results, we found that most of the errors can be classified and adjusted with some phonetic rules. We implement an algorithm to post-process the output label files with these rules. The adjusted results can be applied to adapt the parameters of the HMMs. To increase the reliability, we can iterative train the models. This procedure is also illustrated in Fig. 1. The input is the speech signal and its transcription net file; the output is the INITIAL/FINAL sequence with the associated position.

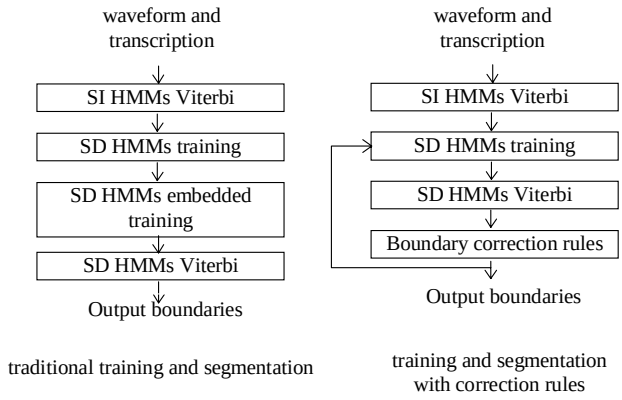

Figure 1. Block diagrams of the two kinds of training and segmentation process

For the traditional training processing, the Speaker Independent (SI) HMMs performs a rough segmentation for the initial training of the Speaker Dependent (SD) HMMs. The parameters of the SD HMMs can be further re-estimated with an embedded Forward-Backward algorithm (HERest). The Viterbi process then outputs the final segmentation with these HMMs. In our case, the re-estimation of HMMs is replaced by an iterative Viterbi, correction and training process. Additional boundary correction rules are applied for the correction. These prior described rules are based on the knowledge from the observations in human correction procedures. This correction is done by searching for the acoustic features matching to the phonetic properties of the units in the local vicinity of the Viterbi alignment boundaries. The features include RMS power, voicing probability and energy bands derived from FFT. The window sizes are varied from 5ms to 20ms according to the features and phonetic types of units. The segmental units are classified into 7 phonetic types, silence, nasal, liquid, fricative, plosive, affricate and vowel. Different rules and features are used for the different combinations of phonetic types, for examples: nasal+vowel, vowel+fricativc, etc. If the case is a sil+plosive, a 5ms window of RMS power will be applied to locate the plosive because there is a short burst of energy when the sound is released. If the specified acoustic features are not found in that area, the boundary is left no change. The adjusted boundaries are further processed to update the parameters of the SD HMMs. These procedures are recursively performed until the average alternation of boundaries is under a threshold.

3.2 Experimental Results

To evaluate the effects of the whole process, a set of manual labeled data is set as the reference. The errors are calculated as the difference between the aligned boundaries and the reference boundaries. The segmentation rate is defined as the percentage of errors within 10ms and 20ms. Without the boundary correction rules, the mean error of the HMMs is 14.2ms, and the segmentation rate is 66.3% (91.2%) within 10ms (20ms). By retraining the HMMs with the boundary correction rules, the average error of the outputs decreases to 8.3ms, and the segmentation rate within 10ms (20ms) increases to 78.4% (96.5%).

4. AUTOMATIC PROSODIC LABELLING

4.1 Multiple Pass Labeling

The prosodic labeling task requires mapping a sequence of feature vectors (mostly derived from the acoustics) to a sequence of labels. The simplest approach to this problem is to treat each feature vector as an independent classification problem, assign a class to it, and move on to the next vector. This approach is too simplistic and fails to represent the inter-relationships between the events being labeled (i.e., to take account of the restrictions that govern the allowable sequences of the events). In the paper of Wightman and Ostendorf[3], these constraints are formularized as a Hidden Markov Model. This is based on a simplified assumption that the current label is only dependent on the previous label. In our experience from manual labeling, the dependency is most on the upper level of units but not the previous label. Since there is a hierarchical structure of prosody, we choose to use a multiple pass top-down procedure for the labeling of break indices. The algorithm contains two principal components. The feature extraction component transforms the various sources of information (the segmental transcription, pitch-tracking results, etc.) into a time-ordered sequence of feature vectors. Thus, if we wish to label break indices on syllabic boundaries, we will need to produce one feature vector for each syllabic boundary. The feature vectors are then classified using decision trees. Decision trees were chosen because they provide a graceful means of handling extremely non-homogeneous features, and the internal structures can be inspected to gain insight into the profit of the features. The multiple pass procedure is very simple. We only spot one kind of break index a time and the sequence is from 5 to 1. The 0 is missing because the reduced syllabic boundary (B0) does not occurred in this read-speech database.

4.2 Prosodic Features

Although there are many features used to determine the break indices, they are all derived from 3 basic acoustic features: duration, energy and F_0. To separate the effect of the segmental intrinsic properties, all these values should be normalized to the z-score values. These features are listed in Table 1. One class of the important features is the temporal information. The information includes the duration of the upper level unit, and the distance of the potential boundary from the beginning or end of the upper level unit. These features are used to take account of the constraints of speech production. We measure these features both in seconds and in number of syllables. The feature F_0 reset should also be particularly mentioned. In order to get a principled representation of the overall shape of the F_0 contour that can suppress the pitch tracking errors. An algorithm used to perform the stylization is based on the technique described in[7]. This algorithm produces a piece-wise linear representation of the F_0 contour. For the boundary between syllable S_1 and S_2, the F_0 reset is defined as f_2-f_1, where f_2 is the F_0 at the beginning of the FINAL part of S_2 and f1 is the F_0 at the end of the FINAL part of S_1. Because the value of the F_0 reset is mostly depending on the lexical tonal combination of the adjacent syllables, these values should also be converted to the z-score according to the tonal patterns.

The features mentioned above are all derived from the acoustic signals. Additional features could be derived from the corresponding text transcription. The location of punctuation mark can be directly copied from the text. It's very useful for the spotting of B4 and B5. The word boundary is another important information that could be derived with a word identification program. Almost all the syllabic boundary inside a word is a B1. We only use these two features because we have no reliable tool for the determination of high level syntactic information.

Symbols	Descriptions	Major Determinant
Eu	end of utterance	B5
Pd	pause duration	B1-B5
Dp	normalized duration of the preceding syllable	B1-B3
Df	normalized duration of the following syllable	B1-B3
Dr	Df / Dp	B1-B3
Ep	normalized energy of the preceding syllable	B1-B3
Ef	normalized energy of the following syllable	B1-B3
Er	Ef / Ep	B1-B3
Ut/Un	total duration or syllables in upper level unit	B1-B5
Bt/Bn	distance (seconds or syllables) from Beginning of upper level unit	B1-B5
Et/En	distance (seconds or syllables) from end of upper level unit	B1-B5
Fr	normalized F_0 reset	B1-B5
Fb	normalized F_0 of the beginning of the following syllable	B4-B5
Fe	normalized F_0 of the ending of the proceeding syllable	B4-B5
Pm	punctuation mark	B4-B5
Wb	word boundary	B1

Table 1. Prosodic features for break indices labeling

4.3 Experimental Results

The experiments are done with the database of one male speaker. 399 paragraphs are used for training and 200 paragraphs for testing. There are 6249 syllables and the same number of labels in the manual labeled test data, because a B5 is labeled at the end of each utterance. Experiment 1 use only acoustic derived features and experiment 2 use all the features. Table 2 and 3 are the confusion matrixes for experiment 1 and 2. The average error rate is 20.3% and 15.1%. We find the confusion can be effectively decreased with the text derived features. This implies the transcriber is affected by the text information and not only use the acoustic information. Although this is unavoidable, we should try to decrease the influences.

Manual Labels	Automatic Labels					Total
	B1	B2	B3	B4	B5	
B1	3211 (83.2%)	581 (15.0)	65 (1.6%)	0 (0%)	0 (0%)	3857
B2	113 (11.1%)	809 (79.4%)	84 (8.2)	7 (0.7%)	5 (0.5%)	1018
B3	14 (2.2)	71 (11.3%)	487 (77.5%)	35 (5.6%)	21 (3.3%)	628
B4	0 (0%)	3 (0.1%)	48 (13.8)	171 (49.0%)	127 (36.3%)	349
B5	0 (0%)	0 (0%)	17 (4.2)	76 (19%)	304 (76.5%)	397
Total	3338	1464	701	289	457	6249

Table 2. Confusion matrix for break indices labeling (without text derived features)

Manual Labels	Automatic Labels					Total
	B1	B2	B3	B4	B5	
B1	3466 (89.8%)	378 (9.8%)	13 (0.3%)	0 (0%)	0 (0%)	3857
B2	98 (9.6%)	832 (81.7%)	78 (7.7%)	8 (0.8%)	2 (0.1%)	1018
B3	15 (2.3%)	65 (10.3%)	510 (81.2%)	25 (4.0%)	13 (2.5%)	628
B4	0 (0%)	4 (1.1%)	34 (9.7%)	186 (53.2%)	125 (35.8%)	349
B5	0 (0%)	0 (0%)	12 (3.0%)	72 (18.1%)	313 (78.8%)	397
Total	3579	1279	647	291	453	6249

Table 3. Confusion matrix for break indices labeling (with text derived features)

5. DISCUSSIONS

In this paper we described the techniques and methodology developed for automatic labeling of segmental and prosodic information for the Mandarin speech database. The results in both the segmental and prosodic labeling are satisfactory in this stage, but there is still a big problem for the prosodic labeling. The experiment is a speaker dependent case. We need large amounts of manual labeled data to train the decision trees. Although most of the features have been normalized, we still have no idea it will work for different speaker or not. More manual labeled data is needed for the further testing. It is also critical that the algorithms should be extended to handle spontaneous speech. All the works have been done so far are focused on read speech. This is because the application is mainly for speech synthesis. The next step is to adapt the models for spontaneous speech and increase the applications for this research.

6. REFERENCES

1. F. Brugnara, D. Falavigna and M. Omologo “Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models”, *Speech Communication 12*, pp. 357-370, 1993

2. Steffen Pauws, Y. Kamp and Lei Willems, “A Hierarchical Method of Automatic Speech Segmentation for Synthesis Applications”, *Speech Communication 19*, pp. 207-220, 1996

3. C. W. Wightman and M. Ostendorf. “Automatic labeling of prosodic patterns”, *IEEE Trans. on Speech and Audio Processing, October*, pp. 469-481, 1994.

4. C. W. Wightman and N. Campbell. “Improved labeling of Prosodic Structure”, *IEEE Trans. on Speech and Audio Processing*, submitted manuscript

5. Chiu-yu Tseng, “A Phonetically Oriented Speech Database for Mandarin Chinese”, *International Congress of Phonetic Sciences*, pp. 326-329, 1995

6. Chiu-yu Tseng and Fu-chiang Chou, “Machine Readable Phonetic Transcription System for Chinese Dialects Spoken in Taiwan”, *Proceedings of the Oriental COCOSDA Workshop*, pp 179-183, 1998

7. P. C. Bagshaw, "An Investigation of Acoustic Events Related to Sentential Stress and Pitch Accents, in English", Proc. 4th. Australian International Conference on Speech Science and Technology. pp. 808-813, 1992