Prosody-dependent Acoustic Modeling for Mandarin Speech Recognition

Tzu-Hsuan Chiu, Chen-Yu Chiang, Yuan-Fu Liao[*]*, Jyh-Her Yang, Yih-Ru Wang and Sin-Horng Chen*

Department of Electrical Engineering, National Chiao Tung University, Taiwan
*Department of Electronic Engineering, National Taipei University of Technology, Taiwan

thchiu7766@gmail.com, gene.cm91g@nctu.edu.tw, yfliao@ntut.edu.tw, neil.yang0204@gmail.com, yrwang@mail.nctu.edu.tw, schen@mail.nctu.edu.tw

Abstract

A study on introducing prosodic information to acoustic modeling (AM) for speech recognition is reported in this paper. It extends the conventional context-dependent (CD) triphone HMM modeling approach to further consider the dependency of phone model on the break type of nearby inter-syllable boundary. Four break types are considered, including major break, minor break, normal non-break, and tightly-coupled non-break. In the training phase, break labeling is automatically accomplished by a Prosody Labeling and Modeling algorithm proposed previously. Then, prosody- and phonetic-dependent phone models are constructed by a standard decision tree-based context clustering of HMMs. The effectiveness of the new AM was examined on a Mandarin syllable recognition task. Experimental results showed that the new approach outperformed the conventional CD-AM on achieving better syllable recognition rate as well as on obtaining a more efficient syllable lattice with better compromise on complexity verse syllable coverage rate.

Index Terms: acoustic modeling, speech recognition, prosody-dependent acoustic model, prosodic break

1. Introduction

Acoustic modeling (AM) in speech recognition (SR) is to build models to represent basic recognition units, such as phones. The most popular approach to AM is the context (phonetic)-dependent (CD) triphone modeling which builds hidden Markov models (HMMs) for phones via considering the coarticulation effects from two neighboring phones. The approach can also be extended to further consider some other affecting factors such as the within-word/syllable and cross-word/syllable dependencies. Motivated by the success of a recent study on using prosodic information to assist in Mandarin SR [1], we propose a prosody- and phonetic-dependent AM in this study aiming at improving the conventional CD triphone modeling via further considering the dependency of phone model on the break type of nearby inter-syllable boundary other than the phonetic effect. It is referred to as PD-AM. The idea is based on our intuition that the acoustic characteristic of the current phone is influenced by the neighboring phones in different degree depending on the break type. The degree is high when there exists no break between them, is low for a minor break with short pause duration, and becomes almost none for major break with long pause duration.

The proposed PD-AM approach can be regarded as an extension of the conventional CD-AM considering both within-syllable and cross-syllable contextual phone dependencies. The modification lies in the finer consideration of the cross-syllable dependency to separate it into four cases for four different break types. With the modification, a more precise control of phonetic contextual influence on triphone modeling can be reached. Moreover, we also let these four break types be engaged with a hierarchical prosodic model (HPM) [2] used in the prosody labeling of the PD-AM training database to build their relations with both prosodic-acoustic features and word-level linguistic features through the HPM. This makes the extra break-type information carried by the trained PD triphone models be useful for helping linguistic and prosodic decoding in further processing. One possible way to realize the idea is to incorporate it into the prosody-assisted Mandarin SR using the HPM [1] to serve as a front end.

Some related works can be found in previous studies. Ostendorf *et al.* [3,4] conducted a pilot study to investigate the effects of prosodic context on AM from a conversational corpus - the Switchboard corpus. However, the recognition performance of the resultant PD-AM could not compete with the state-of-the-art AM due to the lack of abundant prosody-labeled speech data. In [5], Ostendorf *et al*. proposed a PD-AM approach to introducing the dynamic pronunciations of *baseform-to-surface-form phone prediction* using prosodic-acoustic features and word-base linguistic features. Only little effects of prosodic-acoustic features on improving the dynamic pronunciation prediction were found. In [6], Chen *et al.* proposed a PD-AM conditioned on the intonational phrase boundary and the pitch accent. Ni *et al.* [7] proposed a prosody-dependent (PD) tonal syllable AM trained from the '863 corpus' labeled with the break/non-break and stress tags by a bootstrapped automatic prosody labeler. In [8], Huang *et al.* proposed a PD-AM based on the variable-parameter hidden Markov model, in which mean vectors of Gaussian mixture models were functions of the prominence score predicted by a support vector regression method given with prosodic-acoustic features extracted from an N-best word list.

The paper is organized as follows. Section 2 presents the proposed PD-AM approach for SR in detail. Experiment to verify the validation of the proposed PD-AM on a Mandarin syllable recognition task is conducted in Section 3. Some conclusions and future works are given in the last section.

2. The proposed PD-AM approach

Fig. 1 shows a block diagram of the training phase of the PD-AM. Like the training of the conventional CD triphone acoustic modeling, it segments the speech database and uses the decision tree-based training algorithm to generate triphone models. The main differences lie in adding an automatic prosody labeling algorithm to determine break types of all inter-syllable boundaries of the training database and extending the standard decision tree-based context clustering of HMMs to incorporate some extra prosody (break)-related questions. We describe the PD-AM training in more detail as follows.

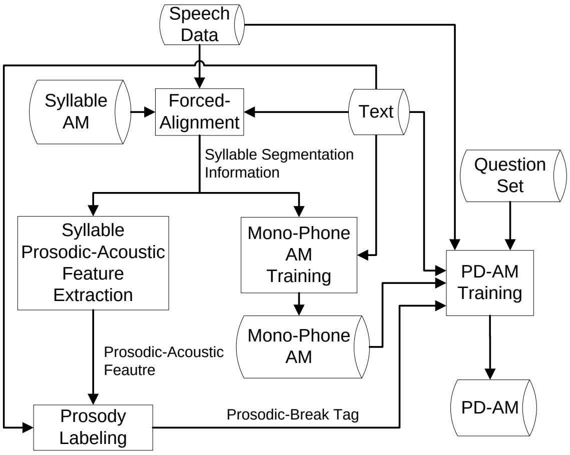

Fig. 1: A block diagram of the PD-AM training

2.1. Experiment database and acoustic features

Throughout the paper, the proposed PD-AM is evaluated on a Mandarin syllable recognition task using a large read speech database TCC300 [9] uttered by 300 speakers, including 150 females and 150 males. A training set containing long paragraphic utterances of 164 speakers (8.3 hours) was used for prosody labeling and acoustic modeling, while a test set of 19 speakers (2 hours) was used for the outside test. Acoustic feature vector used in this study is a 38-dimensional vector composed of 12 MFCC parameters analyzed at a 10-msec frame rate with a 32-msec Hamming window size, their first and second order time derivatives, energy parameter's first and second order time derivatives. Cepstrum Mean Normalization (CMN) is employed to compensate the bias of channel and/or speaker.

2.2. Prosody labeling of the speech corpus

For prosody labeling, an indirect representation of inter-syllable boundary is adopted in this study. All inter-syllable boundaries are classified into four break types: major break, minor break, normal non-break, and tightly-coupled non-break. They are denoted as *BT*0, *BT*1, *BT*2, and *BT*3. For Mandarin speech, *BT*0 and *BT*1 represent respectively non-breaks of reduced syllable boundary and of normal syllable boundary, which have no identifiable pauses between syllables. Second, *BT*2 represents perceivable minor-break boundary with a short pause. Lastly, *BT*3 represents perceivable major-break boundary with a clear long pause.

An unsupervised prosody labeling and modeling algorithm (PLM) proposed previously [2] is employed to perform the prosody labeling. The PLM first classifies all inter-syllable boundaries of the speech database into seven finer break classes {*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4} automatically, and then combines them into four classes via setting *BT*0={*B*0}, *BT*1={*B*1, *B*2-1, *B*2-3}, *BT*2={*B*2-2}, and *BT*3={*B*3, *B*4}.

2.3. Training of PD-AM

The original phonetic question set for CD-AM is then modified to add eight prosody-related questions for the training of PD-AM. They are in the form of "Is it a phone left/right neighboring to a syllable boundary of *BT*0/*BT*1/*BT*2/*BT*3?". Lastly, prosody- and phonetic-dependent phone HMM models (i.e., PD-AM) are constructed by a decision tree-based context clustering of HMMs. For comparison, conventional CD triphone models (referred to as PI-AM) are also generated. There are in total 114 trees (38 phones x 3 states) generated.

It is noted that a silence model and 4 PD short pause (PD-sp) models are also trained. The silence model is of 3 states and used for modeling the long silences existing at the beginning and ending parts of all utterances. The 4 PD-sp models are designed to match the durational characteristics of their corresponding break types. Their topologies are displayed in Fig. 2. For *BT*0, a one-state model with state skipping and non-recurring is adopted to meet its distinct property of non-pause or very short pause. For *BT*1, a one-state model with state skipping and recurring is adopted to match its property of non-pause or short pause. For *BT*2, a three-state model with center state self-recurring and skipping is adopted to fit its property of short pause. For *BT*3, a three-state model with state recurring is adopted to fit its property of long pause.

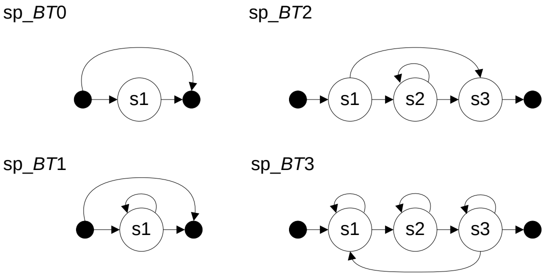

Fig. 2: The topologies of 4 PD short pause models. Note that black solid circle nodes represent NULL states.

Since more information is used in PD-AM, we let its trees grow deeper to have more leaf nodes while setting the average mixture number per leaf node be higher for PI-AM for fair comparison. The total leaf node number and mixture number are 1849 (1595) and 29584 (28710) for PD-AM (PI-AM), respectively. Fig. 3 displays the average log-likelihood evolutions for the training phases of PI-AM and PD-AM. The operations and settings in various stages of these two AM training procedures are described as follow:

Stage 1: Initialization with CI phone models, 1 mixture
Stage 2: Expanding CI-AM to initialize PD-AM/PI-AM, 1 mixture
Stage 3: Tree growing for PD-AM/PI-AM with state-tying, 1 mixture
Stage 4: Increasing mixtures of PD-AM/PI-AM with re-segmentation, 2 mixtures
Stage 5: Increasing mixtures of PD-AM/PI-AM with re-segmentation, 4 mixtures
Stage 6: Increasing mixtures of PD-AM/PI-AM with re-segmentation, 8 mixtures
Stage 7: Increasing mixtures and with re-segmentation, 16 mixtures for PD-AM/18 mixtures for PI-AM.

It can be found from Fig. 3 that PD-AM performs better than PI-AM. To further understand the effects of prosody-related questions on model training, we analyze the locations of prosody-related questions in the 114 trees generated. Table 1 lists the number of trees that the first prosody-related question appears at the top 3 layers. As shown in the table, 63% of trees have the first prosody-related question occurring within the first 3 layers. Note that the average number of layers per tree is 7.032. This shows that prosody-related questions are essential factors to control the tree growing process. We also find that there are in total 230 times of the *BT*3-related questions. Among them, only 14 have phonetic questions of the same side (left or right) appearing in the lower layers of the same tree. This conforms to our intuition that the influence from a neighboring phone becomes almost none for major break with long pause duration.

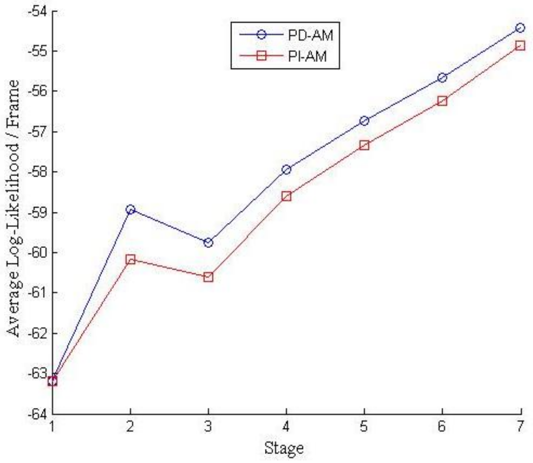

Fig. 3: The average log-likelihood evolution of the PD-AM training phase.

Table 1: Number of trees that the first prosody-related question appears at the *n*-th layer.

layer	1	2	3	> 3
no. of trees	14	32	26	42
cumulative no. of trees	14	46	72	114

3. Syllable recognition using PD-AM

Fig. 4 depicts a block diagram of Mandarin syllable recognition using PD-AM. The task is to recognize the best syllable sequence for the input speech, or to generate a syllable lattice from the input speech for further processing to recognize the best word sequence. It employs the PD-AM with a base-syllable lexicon to generate the output syllable sequence/lattice under the constraint of a grammar.

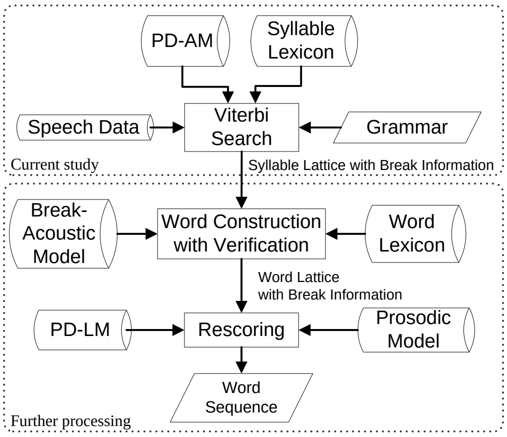

Fig. 4: A block diagram of syllable recognition using PD-AM.

The base-syllable lexicon defines the constituent phone sequence for each of 411 Mandarin base-syllables. Table 2 illustrates an example of entries of 32 break-dependent base-syllables expanded from two prosody-independent base-syllables.

Table 2: An example of the base-syllable lexicon

Lexicon Entry (*BT_SYL_BT*)	P1	P2	P3	P4
*BT*0_yin_*BT*0	*BT*0_yi	e	en_*BT*0	
*BT*0_yin_*BT*1	*BT*0_yi	e	en_*BT*1	
		⋮		
*BT*3_yin_*BT*3	*BT*3_yi	e	en_*BT*3	
*BT*0_bian_*BT*0	*BT*0_b	yi	a	en_*BT*0
*BT*0_bian_*BT*1	*BT*0_b	yi	a	en_*BT*1
		⋮		
*BT*3_bian_*BT*3	*BT*3_b	yi	a	en_*BT*3
		⋮		

Fig. 5 depicts the grammar used in the syllable recognition task. The grammar is a modification of the free syllable grammar by adding four break-dependent short pause models (i.e., PD-sp) shown in Fig. 2 to more strictly confine syllable transitions.

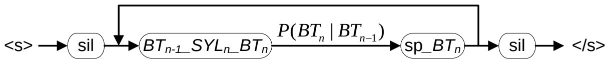

Fig.5: The grammar used in the Mandarin syllable recognition.

Fig. 6 illustrates an example of using PD phone models to form a candidate for recognizing the word “yu-yin-bian-ren” (語音辨認, speech recognition) with inter-syllable break type sequence “*BT*1, *BT*2, *BT*1, *BT*3”.

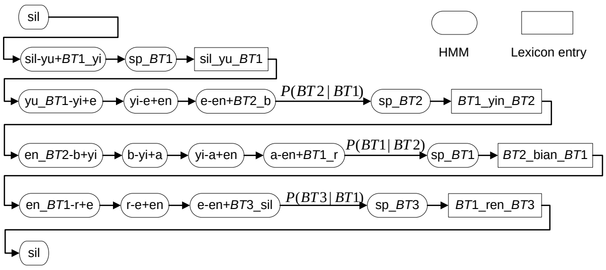

Fig.6: An example of using PD phone models to form a candidate for recognizing the word “yu-yin-bian-ren”. Phone+break-type sequence is: “sil yu *BT*1 yi e en *BT*2 b yi a en *BT*1 r e en *BT*3 sil”.

3.1. Experimental results

We now examined the effectiveness of PD-AM and PI-AM on the syllable recognition task using the TCC300 database. Experimental results are listed in Table 3. As shown in the table, syllable recognition rates of 68.20% and 67.55% were obtained by the PD-AM and PI-AM, respectively. This shows that PD-AM slightly outperformed PI-AM.

Table 3: Experimental results of syllable recognition

	hit	sub.	ins.	del.	total	recognition rate
PI-AM	18235	7889	354	348	26472	67.55%
PD-AM	18364	7790	311	318	26472	68.20%

We also examined the performances of the syllable lattices generated by the PD-AM and PI-AM. Several syllable lattices of different size were generated. Fig. 7 displays the compromises of coverage rate verse lattice size (number of arc) for those lattices. As seen in the figure, PD-AM performed slightly better than PI-AM

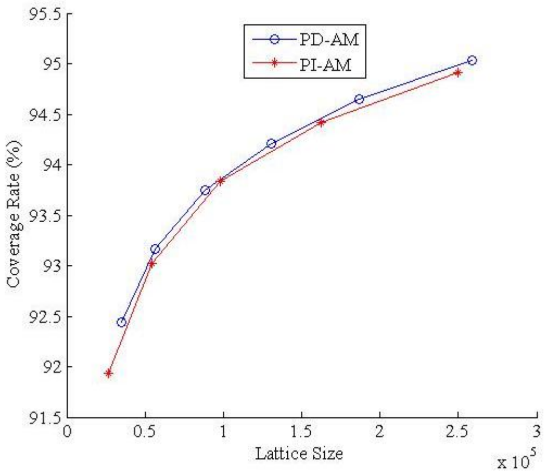

Fig.7: The compromises of syllable coverage rate verse lattice size (number of arc) for the lattices of PD-AM and PI-AM.

3.2. Further processing

Using the syllable lattice generated by the current study, further processing shown in Fig. 4 can be performed in the future to complete the word recognition. First, candidate words can be formed from the syllable lattice by using a word lexicon. Verification of these candidate words can be done using the break-type information carried in the syllable lattice and a break-acoustic model describing the relationship of break-acoustic features, such as pause duration, pitch jump and pre-boundary syllable lengthening, with break type of syllable in various word locations. The purpose of word verification is to exclude some improper word candidates for making the word lattice more compact. Lastly, rescoring of the word lattice can be performed using a prosody-dependent language model (PD-LM) and a prosodic model, such as HPM, to find out the best word sequence.

4. Conclusions

We have presented a new acoustic modeling approach to training prosody- and phonetic-dependent triphone HMM models and shown its effectiveness on Mandarin syllable recognition. Experimental results confirmed that the proposed PD-AM approach outperformed the conventional CD-AM. An extension of the current work to further processing of the syllable lattice for Mandarin word recognition is worth doing in the future. The extra break-type information carried by the PD-AM can be used to help SR in further processing. It is also worth further studying to change the current ML training to a discriminative training using the minimum classification error (MCE) criterion or the maximum mutual information (MMI) criterion.

Acknowledgements

This work was supported in part by the NSC of Taiwan under Contract No. NSC98-2221-E-009-075-MY3 and NSC99-2221-E-009-009-MY3; and in part by the MoE ATU plan. The authors would like to thank the ACLCLP for providing the TCC300 Corpus.

References

[1] Jyh-Her Yang, Ming-Chieh Liu, Hao-Hsiang Chang, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen, “Enriching mandarin speech recognition by incorporating a hierarchical prosody model,” in Proc. ICASSP 2011, Prague, Czech, May, 2011, pp 5052-5055.

[2] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, “Unsupervised joint prosody labeling and modeling for Mandarin speech,” Journal of the Acoustic Society of America, vol. 125, no. 2, pp.1164-1183, Feb. 2009.

[3] I. Shafran, M. Ostendorf and R. Wright, “Prosody and phonetic variability: Lessons learned from acoustic model clustering”, in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 127-131, 2001.

[4] M. Ostendorf et al., “A prosodically labeled database of spontaneous speech,” Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 119-121, 2001.

[5] M. Ostendorf, I. Shafran, and R. Bates, “Prosody models for conversational speech recognition,” in Proc. 2nd Plenary Meeting Symp. Prosody and Speech Process 2003, pp. 147-154.

[6] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, “Prosody dependent speech recognition on radio news corpus of American English,” IEEE Trans. on Audio, Speech and Language Processing, vol. 14 no.1, pp.232-245, January 2006.

[7] C. Ni, W. Liu, and B. Xu, “Using prosody to improve Mandarin automatic speech recognition,” in Proc. INTERSPEECH 2010, Makuhari, Japan, Sept. , pp 2690-2693.

[8] Jui-Ting Huang, Po-Sen Huang, Yoonsook Mo, Mark Hasegawa-Johnson, Jennifer Cole, "Prosody-Dependent Acoustic Modeling Using Variable-Parameter Hidden Markov Models," in Proc. Speech Prosody 2010, Chicago, USA, Apr.

[9] Mandarin microphone speech corpus – TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu.