Prosody Dependent Mandarin Speech Recognition

Chong-Jia Ni，Wen-Ju Liu and Bo Xu

***Abstract*—In this paper, we discuss how to model and train Mandarin prosody dependent acoustic model based on automatic prosody annotation corpus. Based on prosody annotation corpus, we first utilize our proposed methods to train prosody dependent and prosody independent tonal syllable model, and then use these models to get the mixed acoustic models. In this paper, we also utilize tone model to improve the correct rate of tonal syllable through revising the tone of the tonal syllable at certain significant level. When compared with the baseline system, the performance of our proposed mixed speech recognition system improves the correct rate of tonal syllable significantly.**

I. Introduction

PROSODY is generally used to describe aspects of a spoken utterance's pronunciation which are not adequately explained by segmental acoustic correlates of sound units (phones). The prosodic information associated with a unit of speech, say, syllable, word, phrase, or clause, influences all the segments of the unit in an utterance. They are also referred to as supra-segment that transcends the properties of local phonetic context. Prosody encoded in the form of intonation, rhythm, and lexical stress patterns of spoken language conveys linguistic and paralinguistic information such as emphasis, intent, attitude and emotion of the speaker. Prosody is also used by speakers to provide cues to the listener and aid in appropriate interpretation of their speech. This facilitates a method to convey the intent of the speaker through meaningful chunking or phrasing of the sentence, and is typically achieved by breaking a long sentence into smaller prosodic phrases. Two key prosodic attributes described above include prominence and phrasing.

Many speech applications can benefit from corpus annotated with prosodic information, such as automatic speech recognition and understanding. In recent years, there have been a large amount of computational works aimed at prosodic modeling for automatic speech recognition and understanding. There are two major approaches based on the way of training phonetic phone models. One requires a large prosody annotated speech database and constructs prosody dependent allophones model [1-2], and the other doesn't require a large prosody annotated speech database and makes prosody cues as "hidden event variable" [3-5]. K. Chen [1] proposed a novel probabilistic framework in which word and phoneme were dependent on prosody in a way that reduces word error rates relative to prosody independent recognizer with comparable parameter count. In his proposed prosody dependent speech recognizer, word and phoneme models are conditioned on two important prosodic variables: the intonational phrase boundary and the pitch accent. The proposed prosody dependent speech recognizer is able to reduce word error rates by up to 11% relative to prosody independent recognizer with comparable parameter count in experiments based on the prosody annotated corpus—Boston University Radio News Corpus. M. Hasegawa-Johnson [2] described the automatic speech recognition systems, which were variants of a core dynamic Bayesian network model. In this model, the key hidden variables are the word, the prosodic tag sequence and the prosody dependent allophones. Statistical models of the interaction among words and prosodic tags are trained by using the prosody annotated corpus—Boston University Radio News Corpus. K. Chen and M. Hasegawa-Johnson both utilized prosody annotated speech corpus to train prosody dependent allophones models. It requires a large prosody annotated speech corpus. It is very expensive and time consuming to annotate prosody manually. Shriberg et al [3-5] have proposed a different approach that makes use of the acoustic prosodic cues without using explicit prosodic labeling. In their approaches, prosodic events are not explicitly modeled. Instead, prosodic cues are conditioned over a set of hidden event variables representing sentence, disfluency, which are correlated with prosodic events. The advantage of their approaches is that it does not require a large prosody annotated speech corpus. There has been little research on prosody dependent Mandarin ASR based on prosody annotated speech corpus. Very few methods about how to model Mandarin prosody dependent acoustic model and how to train Mandarin prosody dependent acoustic model have been reported. In this paper, we will discuss these issues.

The rest of the paper is organized as follows. In section 2, the prosody dependent speech recognition system is introduced. In section 3, the tone model is presented. The experiment results and analysis are reported in section 4. The conclusion is drawn finally.

II. Prosody Dependent Speech Recognition System

In this section, we first introduce the importance of fundamental frequency in Mandarin speech recognition from

Manuscript received January 25, 2011. This work was supported in part by the China National Nature Science Foundation (No.60675026, No.90820303 and No.90820011), 863 China National High Technology Development Project (No.20060101Z4073, No.2006AA01Z194), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

Chong-Jia Ni is PhD student. His research interests include prosody model, automatic speech recognition and machine learning. (Phone: (+86)10-62659279, e-mail: cjni@nlpr.ia.ac.cn).

Wen-Ju Liu, was with the Electrical Engineering Department Institute of Automation, Tsinghua University , Beijing China. He is now with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing China. (e-mail: LWJ@nlpria.ac.cn).

Bo Xu, was with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing China. Now he is still with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (e-mail: xubo@nlpr.ia.ac.cn).

978-1-4244-9636-5/11/$26.00 ©2011 IEEE

some useful statistic information by utilizing the recognition results of Mandarin speech recognition system, and then multi-space probability distribution hidden Markov model (MSD-HMM), lastly give the prosody dependent allophones based on MSD-HMM.

A. Error rate statistics

In this section, we use the baseline speech recognition system to recognize the speeches used to train the system in order to explain the importance of fundamental frequency information in Mandarin speech recognition. The descriptions about the baseline speech recognition system and the features used in speech recognition are introduced in the section 4. Table 1 lists the recognition results of tonal syllable.

TABLE 1. THE TONAL SYLLABLE RECOGNITION RESULTS OF THE BASELINE SYSTEM

	Cor.	Sub.	Del.	Ins.	Err.
Baseline	88.20	10.71	1.09	0.33	12.13

From Table 1, we can find that the substitute error is the primary error of all the errors, which takes up 88.29% of all. In the substitute error, the initial substitute error is about 22.39%, the final substitute error 24.71% and the tonal substitute error 66.68%. So we can find the decrease of the tonal error rate is important to the decrease of the total error. So, the fundamental frequency information, as input feature, is often used in Mandarin speech recognition. Since fundamental frequency values are not defined in unvoiced region, a commonly used approach to cope with discontinuous fundamental frequency trajectory is to interpolate fundamental frequency in unvoiced segments in order to use fundamental frequency cues in speech applications. This is obviously incorrect. Multi-space probability distribution HMM (MSD-HMM) proposed by K.Tokuda can solve this problem, and has been successfully applied to speech synthesis, speech recognition [6, 7].

B. MSD-HMM

Multi-space probability distribution (MSD) was first proposed by Tokuda et al. to model stochastically the piece-wise continuous fundamental frequency trajectory and was applied to HMM-based speech synthesis [6, 7].

It assumes that the observation space Ω of an event is made up of G sub-spaces Ω_g .That is,

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{1}$$

where Ω_g is a n_g dimensional real space R^{n_g} ,and specified by space index g . Each sub-space Ω_g has its prior probability $p(\Omega_g)$ and $\sum_{g=1}^{G} p(\Omega_g) = 1$. If $n_g > 0$, each space has a probability density function (pdf). If $n_g = 0$, the space Ω_g contains only one sample point.

An observation o , consists of a set of space indices I and a random variable $x \in R^n$, and it is randomly distributed in each sub-space according to an underlying pdf $p_g(V(o))$, where $V(o) = x$. The dimensionality of the observation vector can be different in different sub-spaces. The observation probability of o is defined by

$$b(o) = \sum_{g \in S(o)} p(\Omega_g)\, p_g(V(o)) \tag{2}$$

where $S(o) = I$. The index set of the sub-spaces I that observation o belongs to is determined by the extracted features x at each time instant of observation. A mixture of K Gaussians can be seen as a special case of MSD. The mixture weight associated with the k-th Gaussian component c_k can be regarded as the prior probability of the k-th sub-space $c_k = p(\Omega_k)$.

From the previous analysis of tone recognition error, we can find that the fundamental frequency is the most relevant feature used in recognizing tonal languages. The fundamental frequency only exists in the voiced region. The fundamental frequency contour curve is not continuous, and is piece-wise continuous. The discontinuity of the fundamental frequency contour curve makes the fundamental frequency modeling difficult.

MSD is effective to model the character of the piece-wise continuous fundamental frequency without resorting to unnecessary assumptions. In the voiced region, the fundamental frequency is regarded as one-dimension observation generated from several Gaussian sub-spaces. In the unvoiced region, the fundamental frequency is regarded as a yes-no indicator-like discrete symbol.

An N-state MSD-HMM λ is specified by initial state probability distribution $\pi = \{\pi_j\}_{j=1}^{N}$, the state transition probability distribution $A = \{a_{ij}\}_{i,j=1}^{N}$, and the state output probability distribution $B = \{b_i(\cdot)\}_{i=1}^{N}$,where

$$b_i(o) = \sum_{g \in S(o)} p_i(\Omega_g)\, p_g^i(V(o)), i = 1, 2, \cdots, N \tag{3}$$

Observation probability of $O = \{o_1, o_2, \cdots, o_T\}$ can be written as

$$\begin{aligned} P(O|\lambda) &= \sum_{all\ q} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \\ &= \sum_{all\ q,l} \prod_{t=1}^{T} a_{q_{t-1} q_t} p_{q_t}(\Omega_{l_t}) p_{l_t}^{q_t}(V(o_t)) \end{aligned} \tag{4}$$

where $q = \{q_1, q_2, \cdots, q_T\}$ is a possible state sequence, $l = \{l_1, l_2, \cdots, l_T\} \in \{S(o_1), S(o_2), \cdots, S(o_T)\}$ is a sequence of space indices which is possible for the observation sequence O ,and $a_{q_0 j}$ denotes π_j . The modified forward and backward algorithms can be utilized the computing of observation probability. The modified Viterbi algorithm can be utilized the decoding. These algorithms are similar to traditional HMM algorithms. We do not introduce these algorithms further. The readers who are interested in these algorithms can refer to the references [6, 7].

C. Prosody dependent speech recognition system

Our previous works described some algorithms about the prosodic break automatic annotation and the stress automatic annotation [8-11]. These automatic prosody annotated algorithms are the combination of different classifiers. Through extracting the acoustic, lexical and syntactic related

features and training the prosody models which include prosodic break model and stress model, based on the prosody annotated corpus, we can detect and annotate the prosodic break type and stress type for each syllable in the other corpus. We have annotated the prosodic break type and stress type for each syllable in the "863" and Intel continuous speech corpora which are used in our experiments. When annotating the prosodic break type and stress type for each syllable on these corpora, we only consider whether the syllable is followed by a prosodic break or not and whether the syllable is stressed or not. We only cite these experiments results and list them in Table 2.

TABLE 2. THE ANNOTATION RESULTS ON PART OF MANDARIN CONTINUOUS SPEECH CORPUS

		Precision (%)	Recall (%)	F-Score
Stress	Unstressed	96.9	94.4	95.7
	Stressed	94.1	96.8	95.4
	Mean	95.6	95.5	95.6
Break	Non-break	91.91	95.17	93.51
	Break	92.64	87.89	90.20
	Mean	92.21	92.19	92.20

After the prosody automatic annotation, the tonal syllable turns into the prosody dependent tonal syllable, for example, the syllable "shang3" is annotated as "shang3_3", which means that the syllable "shang3" is stressed and followed by a prosodic break.

After we make the prosody automatic annotation, the number of the prosody dependent allophones multiplies. In our experiments, the numbers of initial and final are 204 before we make the prosody automatic annotation. But after that, the numbers of prosody dependent allophones are 677. So if we still train prosody dependent tri-phone as the tradition, the number of prosody dependent tri-phone is very large. In our experiments, the number of tri-phone in the baseline is 268212, but the number of prosody dependent tri-phone in our prosody dependent speech recognition system is 3071117, which is 11.5 times of the baseline. The speed of decoding is hard to bear. So we proposed the following training acoustic model method.

Step1: Training prosody independent tonal syllable and prosody dependent tonal syllable acoustic model. For prosody independent tonal syllable or prosody dependent tonal syllable, we separately train these models as follows.

Firstly, we train phone or prosody dependent phone model. The method of training the prosody dependent phone model is similar to the method in training the prosody independent phone.

Then, we train prosody independent di-phone or prosody dependent di-phone models. We also train the cross-word di-phone model just as we train the tri-phone HMM model. In order to share data and avoid the problem caused by data sparse, we also make stream-dependent state tying and use decision tree to predict unseen di-phones.

Finally, we utilize dictionary to synthesize di-phones, and get prosody independent tonal syllable or prosody dependent tonal syllable models.

Step2: For prosody independent tonal syllable and prosody dependent tonal syllable models, we compose the prosody independent and prosody dependent mixed syllable models.

In our experiments, the number of prosody independent tonal syllable is 1302, and the number of prosody dependent tonal syllable is 4523. So after combining prosody independent tonal syllable model with prosody dependent tonal syllable model, the number of prosody independent tonal syllable model and prosody dependent tonal syllable model is 5825. In our experiments, the number of models in the "TriPho-Pro-MSD-HMM" speech recognition system which is constructed by utilizing the traditional method is 3071117, but in our proposed mixed speech recognition system "Syl-Syn-Pro-MSD-HMM", the number of models is 5825. The number of models decreases greatly. In this paper, we use MSD-HMM to model prosody independent tonal syllable and prosody dependent tonal syllable models.

III. TONE MODEL

In section 2.1, we have introduced that the fundamental frequency information is important to Mandarin speech recognition. Therefore, there are lots of works to model tone model and use tone model to rescore the results of speech recognition system at lattice or n-best in order to improve the performance of speech recognition system. In this paper, we also use tone model to improve the correct rate of tonal syllable, but the way of using tone model does not same with the traditional methods.

First, the fundamental frequency values of a given continuous speech utterance are computed by using the RAPT algorithm [12], and the speech utterance is aligned by using LVCSR system in order to acquire the accurate syllable boundary.

Second, the fundamental frequency contour of each syllable is divided evenly into three sections, and the average logarithmic fundamental frequency value of each section is computed. Therefore, for each syllable, we can extract three features. Considering the influence of the context, we also compute features on the previous and the following syllable of the current. The computing method is similar as that of on the current syllable.

Finally, we also compute the "3-gram" probability features of each syllable in the contextual window, which consists of the previous and the following syllable.

Therefore, the total number of the features used in our experiments is 12. We use multi-layer perceptron (mlp) to train the tone model. When testing on "863" testing set, the 83.88% correct rate of tone could be acquired.

We directly use the tone model to revise the tone of tonal syllable at certain significant level f . Let us use p mean the probability score of tone model of the current tonal syllable. If $p > f$, we will revise the tone of the tonal syllable; if not, we will not do that . This method can also be used in the decoding process directly.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. The corpora and experiments setup

The data corpora applied in the experiments are provided by Chinese National Hi-Tech Project 863 and Intel for Mandarin large vocabulary continuous speech recognition

(LVCSR) system development [13]. In "863" speech corpus, 83 male speakers' data are employed for training (48373 sentences, 55.6 hours) and 6 male speakers' for testing (240 sentences, 17.1 minutes). In Intel speech corpus, Intel Beijing-Accent Mandarin corpus is used. So the time of speeches used to train acoustic model is about 110 hours.

Table 3. The simple descriptions about the six speech recognition systems.

System	Observation dimension	Stream Number	State Number
TriPho-HMM (Baseline)	39	1 stream	5 state 3 emitting
TriPho-MSD-HMM	42	2 stream	5 state 3 emitting
Syl-MSD-HMM	42	2 stream	8 state 6 emitting
Syl-Pro-MSD-HMM	42	2 stream	8 state 6 emitting
TriPho-Pro-MSD-HMM	42	2 stream	5 state 3 emitting
Syl-Syn-Pro-MSD-HMM	42	2 stream	8 state 6 emitting

Table 3 gives the simple descriptions about the six speech recognition systems. In Table 3, "TriPho" means the cross-word tri-phone (initial and final) is the object of modeling; "Syl" means the tonal syllable is the object of modeling and the system is obtained by using Step1 only in section 2.3. "Syl-Syn" means we use Step1 and Step2 in our proposed method to train and get the system. "Pro" means the prosody dependent speech recognition system. "HMM" means that we use HTK tools to train and get the HMM models [14]. "MSD-HMM" means that we use HTS tools to train and get the MSD-HMM models [15]. The dimension of observation is 39, which means the 12 MFCC and 1 energy and their one-order and two-order difference. The dimension of observation is 42, which means 39 MFCC and fundamental frequency and its one-order and two-order difference. In 1-stream system, 39 MFCC is one stream. In 2-stream system, 39 MFCC is one stream, fundamental frequency and its one-order and two-order difference is one stream. In all 2-stream systems, we design two question sets respectively and use decision tree to make stream-dependent state tying. The structure of HMM or MSD-HMM in the "TriPho" systems has 5 states from left to right, 3 emitting distributions and no state skipping, except "sp" (short pause) model with 3 states, 1 emitting distribution. The structure of HMM or MSD-HMM in the "Syl" systems has 8 states from left to right, 6 emitting distributions and no state skipping. Each emitting distribution is modeled by 16 Gaussian mixtures.

B. Experimental results and analysis

In this section, we will list six speech recognition systems experimental results on 863-Test set. From the approach which we get the models, we all know that in 863-Test set, the models we have gotten are speaker-independent. Table 4 lists the experimental results about tonal syllable.

First, from Table 4, we can find that these systems with fundamental frequency information as input features can improve the correct rate and lower the error rate.

Second, prosody information can improve the performance of speech recognition. By comparing "Syl-MSD-HMM" system with "Syl-Pro-MSD-HMM" system or "TriPho-MSD-HMM" system with "TriPho-Pro-MSD-HMM" system, it is obvious that prosody dependent speech recognition systems are superior to prosody independent speech recognition systems.

Third, the performance of "Syl-MSD-HMM" and "Syl-Pro-MSD-HMM" system is poorer when compared with the "TriPho-MSD-HMM" system because of the decrease of the number of model. After adding the number of the model, the system "Syl-Syn-Pro-MSD-HMM" achieves better experimental results.

TABLE 4. THE EXPERIMENTAL RESULTS ON 863-TEST SET.

System	Corr.%	Sub.%	Del.%	Ins.%	Err.%
TriPho-HMM (Baseline)	61.88	37.58	0.54	0.51	38.63
TriPho-MSD-HMM	75.58	23.72	0.70	1.53	25.95
Syl-MSD-HMM	72.18	27.25	0.57	0.32	28.14
Syl-Pro-MSD-HMM	75.39	24.10	0.51	0.38	24.99
TriPho-Pro-MSD-HMM	76.25	22.48	1.27	1.05	24.80
Syl-Syn-Pro-MSD-HMM	76.18	23.28	0.54	0.32	24.13

Now, we use the tone model to revise the results of speech recognition. In our experiments, when significant level f is 0.8, there is a better result. Table 5 lists the experimental results.

TABLE 5. THE EXPERIMENTAL RESULTS ON 863-TEST SET AFTER USING TONE MODEL.

System	Corr.%	Sub.%	Del.%	Ins.%	Err.%
Syl-Syn-Pro-MSD-HMM + Tone model	76.31	23.15	0.54	0.32	24.01

V. CONCLUSION AND DISCUSSION

In this paper, we introduce the acoustic processing and modeling of the supra-segmental characteristics of speech. We model fundamental frequency by utilizing MSD, and use automatic prosody annotated methods to annotate syllable prosodic break type and stress type in continuous speech corpus. We also utilize different methods to train prosody dependent tonal syllable model in order to overcome the data sparse problem after prosody annotation. At the same time, we compose the prosody independent tonal syllable model and prosody dependent tonal syllable model in order to overcome the problem of insufficiency of model number. After decoding, we use tone model to revise the results of the speech recognition system, which can also improve the correct rate. Of course, we should be clear that the large speech corpus is required when training the prosody dependent allophone models. In the future, we will introduce the prosody dependent language model and prosody dependent dictionary when decoding, and obtain Chinese characters.

REFERENCES

[1] K. Chen M. Hasegawa-Johnson, A. Cohen, S. Borys, Sung-Suk Kim, J. Cole and Jeung-Yoon Choi, "Prosody Dependent Speech Recognition on Radio News Corpus of American English", IEEE Trans. Audio, Speech and Langue Processing, 14(1): 232-244, 2006.

[2] M. Hasegawa-Johnson, K. Chen, J. Cole S. Borys, Sung-Suk Kim, A. Cohen, T. Zhang etc., “Simultaneous Recognition of Words and Prosody in Boston University Radio Speech Corpus”, Speech Communication, Vol. 46, 418-439, 2005.
[3] E. Shriberg and A. Stolcke, “Direct Modeling of Prosody: An Overview of Application in Automatic Speech Recognition”, in Proc. ISCA Int. Conf. Speech Prosody, Nara, Japan, 2004.
[4] D. Vergyri, A. Stolcke, V. Gadde, L. Ferrer and E. Shriberg, “Prosodic Knowledge Sources for Automatic Speech Recognition”, in Proc. ICASSP, 2003.
[5] A. Stockle, E. Shriberg, D. Hakkani-Tur and G. Tur, “Modeling the Prosody of Hidden Events for Improved Word Recognition”, in Proc. EuroSpeech, 1999.
[6] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, “Multi-Space Probability of Distribution of HMM”, IEICE Trans. Inf. and Sys. Vol. E85-D, No.3, March 2002.
[7] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling”, in Proc. ICASSP, 1999.
[8] Chong-Jia Ni, Wen-Ju Liu and Bo Xu, “Automatic Prosody Boundary Labeling of Mandarin Using Text and Acoustic Information”, in Proc. ISCSLP, Kunming, China, 2008.
[9] Chong-Jia Ni, Wen-Ju Liu and Bo Xu, “Mandarin Stress Detection Using Hierarchical Model Based Boosting classification and regression tree”, in Proc. IJCNN, 2010.
[10] Chong-Jia Ni, Wen-Ju Liu and Bo Xu, “Mandarin Automatic Prosodic Break Detection and Annotation Based on Complementary Model,” submitted to EURASIP Journal on Audio, Speech and Music Processing.
[11] Chong-Jia Ni, Wen-Ju Liu and Bo Xu, “Mandarin stress detection based on complementary model,” submitted to Computer Speech and Language.
[12] Talkin,A. D. “A Robust Algorithm for Pitch Tacking (RAPT)” in Speech Coding and Synthesis, edited by W. B. Kleijn and K.K. Paliwal.
[13] S. Gao, et al, “Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR”, in Proc. ICSLP, 2000
[14] The HTK 3.4.1 Toolkit, Online, http://htk.eng.cam.ac.uk/
[15] HTS 2.1, Online: http://hts.sp.nitech.ac.jp/, accessed on 2009.