Detection-Based Accented Speech Recognition Using Articulatory Features

Chao Zhang #1, Yi Liu #2, Chin-Hui Lee *3

Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Science and Technology
Tsinghua University, Beijing, China
[1] zhangc@cslt.riit.tsinghua.edu.cn, [2] eeyliu@tsinghua.edu.cn

* *School of Electrical and Computer Engineering*
Georgia Institute of Technology, Atlanta, USA
[3] chl@ece.gatech.edu

***Abstract*— We propose an attribute-based approach to accented speech recognition based on automatic speech attribute transcription with high efficiency detection of articulatory features. In order to utilize appropriate and extensible phonetic and linguistic knowledge, conditional random field (CRF) is designed to take frame-level inputs with binary feature functions. The use of CRF with merely the state features to generate probabilistic phone lattices is then utilized to solve the phone under-generation problem. Finally an attribute discrimination module is incorporated to handle a diversity of accent changes without retraining any model, leading to flexible "plug 'n' play" modular design. The effectiveness of the proposed approach is evaluated on three typical Chinese accents, namely Guanhua, Yue and Wu. Our method yields a significant absolute phone recognition accuracy improvement 5.04%, 4.68% and 6.06% for the corresponding three accent types over a conventional monophone HMM system. Compared to a context-dependent triphone HMM system, we achieve comparable phone accuracies at only less than 20% of the computation cost. In addition, our proposed method is equally applicable to speaker-independent systems handling multiple accents.**

I. Introduction

Most state-of-the-art automatic speech recognition (ASR) systems fail to perform well when the speaker has a regional accent different from that of the standard language the systems were trained on. Accent is a crucial bottleneck for an extensive usage of speech-enabled applications across a large population in China since all Chinese speakers share the same ideographic Chinese characters but with different pronunciations due to regional accents. There are seven major dialects in China: Guanhua, Yue, Wu, Xiang, Gan, Min, and Kejia [1]. Linguists often regard each of them as a distinct language [2]. Statistics indicate that about 80% of Putonghua (the standard Chinese) speakers have regional accents, and 44% among them have heavy accents [3].

Conventional methods handle impact of accent by focusing on pronunciation modelling for acoustic or phonetic accent changes at different levels [2, 4–6]. Augmented pronunciation dictionary and phone set extension are commonly used methods for modelling phonetic changes [4]. Maximum a posteriori (MAP) [7] and maximum likelihood linear regression (MLLR) [8] are always applied to adapt the standard models to fit acoustic characteristics of certain accents in order to cover acoustic changes [5]. State-level pronunciation modelling and acoustic model reconstruction can be used to cover accent variations without degrading the performance on standard or other accented speech [2, 6]. However, it is known that the variations within accent changes are often complicated covering both complete and partial changes [2], a diversity of articulation changes and shifts are important features for accent variations. Hence, a method of incorporating articulation information is needed when attempting to handle precise accent variations.

Recently, an automatic speech attribute transcription (ASAT) paradigm was proposed to provide additional information to ASR through the integration of acoustic and phonetic knowledge [9, 10]. ASAT is a bottom-up approach based on a set of detected speech attributes and contains three major components: (1) a bank of attribute detectors to spot acoustic cues from continuous speech; (2) a multi-level event merger that combines attributes into phone and other higher-level units; and (3) an evidence verifier that validates the recognition decisions [10].

Previous work showed that the ASAT strategy is able to take the advantage of linguistic information described by features and rules for robust model generation [11, 12]. Since accent variations have proven at different levels, and articulated changes and shifts (e.g., aspirated, nasal, voiced, etc.) have great impact on the recognition accuracy, the combination of articulatory features with detectors in ASAT should have a strong ability to model the diversity of accent changes, leading to improved recognition performance.

In this paper, we propose an ASAT-based approach to accented speech recognition with high efficiency detection of articulatory features. Compared to the conventional HMM-based ASR systems, the use of articulatory features associated with context-dependent HMM detectors can incorporate more linguistic knowledge, and is able to capture a diversity of coarticulation effects within accented speech. In order to utilize suitable and extensible phonetic and linguistic knowledge, conditional random field (CRF) is designed as frame-level input and with binary features. Meanwhile, we suggest the use of CRF with merely state features to generate probabilistic phone lattice instead of the commonly used one-best phone hypothesis to solve the phone under-generation problem. It achieves a good balance of insertion and deletion in evidence verification, leading to efficient feature processing at high levels. The experimental results show that our proposed approach achieves comparable recognition accuracy at a much faster recognition speed when compared to triphone HMM ASR systems. Furthermore, we also propose an attribute discrimination module that handles accent changes without retraining any model in the system leading to improved flexibility of the system. Finally our proposed approach covers more accent variations in different accents, leading to an enhanced performance for multi-accent speech recognition tasks.

The rest of the paper is organized as follows. In Section II, we describe the speech attributes as well as three typical accents – Guanhua, Yue and Wu in Chinese used in this paper. In Section III, we introduce our proposed detection-based ASR using articulatory

TABLE I
PUTONGHUA PHONE LIST IN TERMS OF ARTICULATORY FEATURES

Category	Attribute	Phone Set
Place	alveolar	d l n t
	bilabial	b m p
	dental	c s z i1
	labiodental	f
	palatal	j q x a o e ei i u v
	palato-alveolar	ch r sh zh i2
	retroflection	er
	velar	g h k ng
Manner	affricative	c ch j q z zh
	fricative	f h r s sh x
	lateral	l
	nasal	m n ng
	stop	b d g k p t
	N/A	ALL_VOWELS
Aspirated	aspirated	c ch k p q t
	unaspirated	b d g j z zh
	N/A	f h l m n r s sh x ng ALL_VOWELS
Voicing	voiced	l m n r ng ALL_VOWELS
	unvoiced	b c ch d f g h j k p q s sh t x z zh
Height	high	i i1 i2 u v
	low	a
	middle high	o e
	middle low	ei er
	N/A	ALL_CONSONANTS
FrontEnd	back	o e u
	central	a er i2
	front	ei i v i1
	N/A	ALL_CONSONANTS
Rounding	rounded	o u v
	unrounded	a e ei er i i1 i2
	N/A	ALL_CONSONANTS

features and other key modules of ASAT. In Section IV, experimental results of using detection based ASAT with articulatory features on three Chinese accents are presented. Finally we summarize our findings in Section V.

II. ATTRIBUTES FOR CHINESE ACCENTED SPEECH

A. Articulatory Features in Putonghua

Chinese is a syllabic language, with each written Chinese character pronounced as one of the 416 non-tonal syllables in Putonghua. A syllable can be decomposed into an initial followed by a final. A pair of initial or final usually corresponds to one to three phones. Initials and finals are commonly used as sub-word units in Putonghua ASR systems. On the other hand, as pointed out in [13], finer-grained units instead of phonemes are more appropriate for modelling pronunciation variations.

Articulatory features, as symbolic indicators used in acoustic phonetics to characterize how phones are produced using related articulators and the airflow from the lungs, can be used to formulate linguistic knowledge for pronunciation changes caused by either regional accent or coarticulation as context-dependent rules associated with substitutions of different features. This enables the incorporation of knowledge into modelling [1, 13, 14]. In this paper, 31 articulatory attributes are chosen, which belong to 7 categories listed in Table I. Four attributes belonging to 'Place', 'Manner', 'Aspirated' and 'Voicing' correspond to minimal units to distinguish consonants, and similarly, one 'Place' attribute together with three attributes belonging to 'Height', 'FrontEnd' and 'Rounding' respectively are minimal units to discriminate vowels [1]. It is remarkable that 'Place' for consonants and vowels have been defined differently in acoustic phonetics [12], we combine the features of the two classes into one. Detail for the set of attributes and the attribute-to-phone conversion rules in Putonghua are listed in Table I. All these mappings are based on consensus in linguistics [1].

B. Guanhua, Yue and Wu Accents

Regional dialect speakers can hardly avoid the influence of their language in the process of second language acquisition, defined as negative transfer in linguistics, which results in difficulties in terms of variations on phoneme inventories, syllable structures, tones, grammar and vocabulary. In this paper, the first type of the variation is emphasized, and we demonstrate why using ASAT with articulatory features is an appropriate and potential recipe for handling accent variations.

In order to evaluate the general effectiveness of our proposed approach, three typical accents are selected in our study – Guanhua, Yue and Wu. These entire three accents cover a speaking population of hundreds of millions and they represent quite different pronunciations in terms of phonology, lexical and syntactic structures [2], for instance, linguists have shown only 60% of Yue is even close to Putonghua [2]. Guanhua is the dialect that Putonghua is based on, whose phoneme inventories are almost the same as Putonghua. In contrast, Yue and Wu have different initials and finals individually. For example: there are no palato-alveolar affricatives/fricatives 'zh', 'ch', 'sh' in Yue and Wu, therefore, 'zh', 'ch', 'sh' are often pronounced as 'z', 'c', 's' in these accents; Yue adopts velar nasal 'ng' as an additional final while it has no retroflection final 'er'; Wu initials have extra voiced consonants that are voiceless in Putonghua, for instance the stop 'd' is voiceless in Putonghua and pronounced as voiced in Wu is an interesting case in this point. As a result, these differences may cause an initial/final be pronounced into a different one when Yue and Wu people speak Putonghua.

Linguistic rules for Chinese accent changes can be naturally explained with articulatory features, therefore, it provides us with an intuitive idea that we can improve Chinese accented speech recognition by covering confusing articulatory features instead of phoneme changes.

III. DETECTION-BASED ASR USING ARTICULATORY FEATURES

Figure 1 illustrates our proposed detection-based system for Chinese accented speech recognition. Following the ASAT paradigm, this system consists of three parts: (1) a bank of speech attribute detectors, (2) an attribute-to-phone merger, and (3) an evidence verifier. Different from previous studies where artificial neural networks (ANNs) were used in [11, 12], we use triphone HMM detectors for high efficiency detection. Instead of neural net features, we use CRF with binary features for the ease of formulating acoustic phonetic rules as state feature functions of CRF [11]. Our CRF has no transition

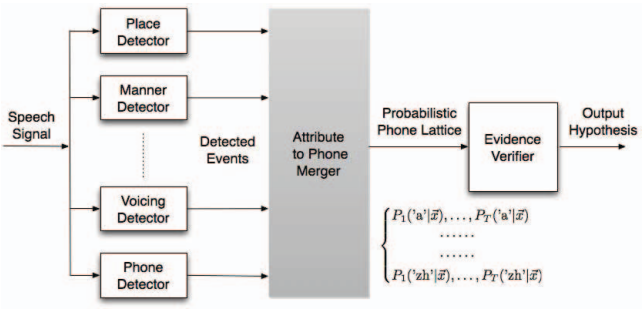

Fig. 1. A block diagram for the detection-based phone recognition system

TABLE II
EVALUATION OF GUANHUA ARTICULATORY ATTRIBUTE DETECTORS

Category	System Correct%		Frame Correct% (Guanhua)		
	Guanhua	Yue	T=0	T=1	T=2
Place	79.54	**68.09**	58.52	62.14	65.10
Manner	84.61	83.30	75.09	79.12	81.57
Aspirated	83.85	82.59	85.19	87.53	89.29
Voicing	88.26	88.20	77.38	81.52	84.01
Height	85.31	83.44	74.93	79.07	81.44
FrontEnd	85.59	83.30	78.03	82.10	84.41
Rounding	85.70	85.12	75.56	79.80	82.77

TABLE III
CORRECTNESS OF 'PLACE' ATTRIBUTES BY GUANHUA DETECTOR

Attribute of Place Category	Segment Correct%	
	Guanhua	Yue
bilabial	81.8	79.0
labiodental	92.2	90.9
dental	82.3	**76.7**
alveolar	75.4	72.3
palato-alveolar	79.3	**71.0**
palatal	80.7	80.2
velar	80.9	77.3
retroflection	90.8	**66.1**

features. Furthermore output probabilistic phone lattices are used to replace the conventional one-best hypotheses. Hence, we attempt to address the phone over-deletion problems in CRF [11, 15].

A. Speech Event Detection

Both articulatory features and phones are used as attributes in our system. We use context-dependent HMMs to detect articulatory features since they are able to capture accent-related coarticulation on articulatory features with high efficiency. Thus, we built a set of triphone HMM detectors for every articulatory feature category, and detectors are used for recognition rather than key word spotting to guarantee that attributes of the same category will not be detected together. Meanwhile, we used monophone HMMs to build phone attribute detector.

In general, Yue and Wu are regarded as far different from Guanhua and Putonghua. Traditional method automatically generated numerous accent variations in terms of phoneme insertions, deletions and substitutions [2, 4, 5]. Modelling for every phoneme change individually to cover accent is a trivial and yet sophisticated task. However, through the use of articulatory attribute detectors, we are able to categorize the phoneme changes into a much smaller number of confusing articulatory attributes as some acoustic phoneticians suggested. We investigated the articulatory attribute confusion in Yue accent as an example. The articulatory attribute detectors whose performances are listed in Table II were trained on 5 hours of Guanhua data. We evaluated the detectors on both the Guanhua and Yue testing sets.

In Table II, "T = 0, 1, 2" means we consider the frames at detected segment boundaries as be correctly detected if they are of the same value as any frame in reference transcription to whom the time difference is less equal than 0, 1 and 2 frames. We study frame-level correctness due to the well-known noisy nature of the segment boundaries.

Table II implies that Yue accent has impacts on the performance of the detectors. We explored the confusion matrix of the Place detector, and listed the correctness of its attributes in Table III. It is clear that bilabial, labiodental and palatal show robustness across the two languages while dental, palato-alveolar and retroflection showed severely degraded performance on the Yue accent. By comparing the confusion matrices of the two accents, we find: (1) the degradation of palato-alveolar was mainly caused by its confusion with dental that coincides with the rule we have mentioned that Yue speakers tend to mispronounce 'zh', 'ch', 'sh' as 'z', 'c', 's'; (2) most of the error samples of retroflection in Yue were misrecognized as either palatal or deleted, since the absence of retroflection vowel 'er' in Yue causes the Yue speakers to pronounce 'er' as its closest vowel 'e' or deleted when 'er' is used as the coda of a syllable; (3) more dental are deleted in the Yue testing set, and we find more deletions of 'z', 'c' and 's' in HMM phone recognition on the same testing set. Hence, accent variations can be represented in a brief and fundamental way in ASR in terms of articulatory attribute confusions.

B. Event Merger

The role of an attribute-to-phone mapping merger is to combine the detected event streams using different weights into a probabilistic phone lattice and deliver it to the evidence verifier. For example, in our study, we expect the retroflection vowel 'er' to be pronounced as the palatal vowel 'e' by Yue speakers. Therefore, the merger should be able to explore such changes from the underlying data according to linguistic rules and learn proper weights for how often and in what context this change happens. A powerful and widely used tool for such purpose is conditional random fields [16].

CRF is an undirected graph model as illustrated in Part (A) of Figure 2. It is capable of integrating redundant input sequence $\vec{x} = \{x_1, x_2, \ldots, x_T\}$ into its most probable label sequence $\vec{y} = \{y_1, y_2, \ldots, y_T\}$ in terms of conditional probability $P(\vec{y}|\vec{x})$ [17]. As an event merger, $\vec{x}$ is the detected event streams grouped by frames, $\vec{y}$ can be the one-best phone hypothesis. CRF can be presented more formally as,

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \vec{x}) \right\}, \qquad (1)$$

where $f(y_t, y_{t-1}, \vec{x})$ is called a feature function, and λ is its weight. $f(y_t, y_{t-1}, \vec{x})$ and λ can be any real number, since the probability meaning of $P(\vec{y}|\vec{x})$ is guaranteed by partition function $Z(\vec{x})$, which is obtained by summing on every hypothesis $\vec{y}^{'}$. $Z(\vec{x})$ is defined as:

$$Z(\vec{x}) = \sum_{\vec{y}^{'}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(y_t^{'}, y_{t-1}^{'}, \vec{x}) \right\}. \qquad (2)$$

Therefore, learning is to find proper weights from the training data for every feature functions maximizing the conditional probability, while decoding is to find the one-best sequence based on the weighted feature functions and the input observations. Detail of training and decoding of CRF can be found in [17].

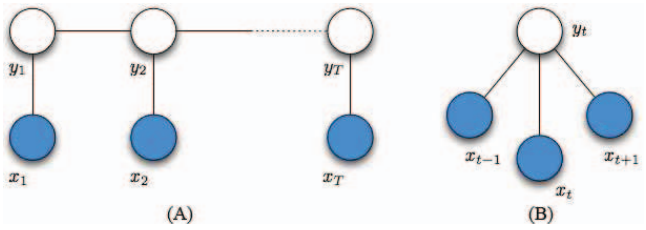

Fig. 2. Graphical representation for CRF and Logistic Regression

One of the key issues in CRF is the design of the feature functions, $f(y_t, y_{t-1}, \vec{x})$, which can be classified into two categories: state and transition feature functions. A state feature function represents a state that some event happens when the output label is a particular value. For example, the feature function for 'er' to be pronounced as a palatal vowel would be

$$s(y, \vec{x}, t) = \begin{cases} 1, & \text{if } y_t = \text{'er' and palatal}(x_t) = true \\ 0, & \text{otherwise} \end{cases} . \quad (3)$$

Transition functions are often defined in a similar manner in ASR that counts for transitions between two labels whose values match the definition of the function [11, 15].

In previous work that uses CRF to combine detected events [11, 15], both state and transition features are used. Since the state features devastatingly reveal that the current frame is of a different label from the previous state and it causes transition features to overwhelmingly point out that the state have changed, numerous deletion errors (several times more than insertions) would be unalterably generated, and is detrimental to system performance [11]. Therefore, our CRF mergers do not use transition functions, and it outputs a probabilistic phone lattice instead of the one-best phone sequence hypothesis, so that the system is able to generate balanced inserted and deleted phones in the evidence verifier. Our state feature functions are described as follows:

1) Presence Features: These features describe the presence of each attribute at current frame.

2) Distinction Features: These features are knowledge-motivated and are employed to present the existence of possible combinations of attributes belonging to 'Place', 'Manner', 'Aspirated' and 'Voicing', and possible combinations of the attributes for 'Place', 'Height', 'FrontEnd' and 'Rounding' at the current frame.

3) Window Features: Every presence feature and every distinction feature at the previous two and the next two frames. Window Features are engaged since we have shown in Table III that values at the surrounding frames do help correcting the detections.

It is remarkable that our usage of CRF without transition features can be regarded as Logistic Regression as has been discussed in detail in [17]. Graphical representation of Logistic Regression is illustrated in Part (B) of Figure 2.

C. Evidence Verifier

Our evidence verifier is similar to [12]. We build a 3-state, left-to-right HMM model for each phone, whose emission probabilities are phone probabilities generated by the event merger. Hence the evidence verification can be accomplished by various decoding techniques. In this study, we build context-independent free grammar decoding network that every phone has the same entrance probability. We use the Viterbi algorithm to search over the network and generate the one-best hypothesis. We can obtain a hypothesized phone string with balanced insertion and deletion errors by setting a suitable word penalty score.

IV. Recognition Experiments

The 863 regional accent speech corpus [18] was used in all experiments to evaluate the effectiveness of our proposed approach. This database is the largest and most commonly used for Chinese accented speech recognition tasks [4, 19]. All data were sampled at 16kHz with a 16-bit precision.

Table IV shows the detailed information of speakers, data durations, and the total phone counts used in all experiments. The HMM topology is three-states, left-to-right without skips. All triphone HMMs for detectors and the baseline systems were built using HTK decision tree based state-tying [20]. The acoustic features were 39-dim vectors with $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. CRF and SVM were trained by CRF++ and LIBSVM toolkit, respectively [21, 22]. All systems were evaluated with free grammar phone recognition rather than syllable or short phrase recognition due to the fact that we evaluated the absolute performance improvement through modelling of accent variations. It is shown in Table IV that each phone has 3,193 samples on average for training. We believe it is sufficient for reliable acoustic model generation.

TABLE IV
Data Sets Separation in Experiments

Data ID	Duration	Phone Number	Speaker Number	Utterance Number	Type
TrainG	4.2h	102,646	60	2,144	Guanhua
TestG	2.9h	70,397	20	1,274	
TrainY	5.1h	105,167	60	2,317	Yue
TestY	2.8h	70,297	20	1,309	
TrainW	5.1h	108,315	60	2,611	Wu
TestW	2.9h	68,173	20	1,283	

TABLE V
Details of Triphone Detectors

Train Set	TrainG	TrainY	TrainW
Attribute	State Number	State Number	State Number
Aspirated	48	46	47
FrontEnd	183	185	182
Height	179	176	176
Manner	147	140	148
Place	230	233	228
Rounding	104	109	106
Voicing	22	22	23

A. Accent-Based Articulatory Attribute Detectors

We built seven triphone articulatory attribute detectors with 6 Gaussian components per state and one 16 Gaussian components per state context-independent phone detector for each accent individually. We listed detail of triphone HMMs detectors in Table V. We constructed a CRF attribute-to-phone merger for each accent. CRF for Guanhua, Yue and Wu accents have 260,370, 272,245 and 272,745 features, respectively. Table VI gives the comparisons between our systems and the HMM baseline systems. Triphone HMM system for Guanhua, Yue and Wu accent has 514, 572 and 574 tied-states, respectively. Each system in Table VI was evaluated on the testing set of the same accent that the system was trained on.

It is known that redundant deletions compared to insertions in using transition features were often generated in previous CRF studies and therefore led to severe performance degradations. Such effects bring additional difficulties for high-level processing, e.g., modelling for pronunciation variations of syllable structures [13]. On the other hand, from Table VI, we can see that our proposed method is able to achieve a good balance of insertion and deletion errors due to the fact that no overwhelming evidences that indicate the label has to retain the same labels generated by transition feature functions, imposed on the phone lattice. Meanwhile, the HMM topology as well as suitable word penalty scores used in the evidence verifier help to balance the inserted and deleted phones.

In addition, our systems perform significantly better than context-independent HMMs that reveals CRF with linguistics-derived binary features is able to capture the underlying distinctions of phonetically

TABLE VI
COMPARISONS OF SYSTEMS TRAINED AND TESTED ON THE SAME ACCENT

System	Accent	Phone Correct%	Phone Accuracy%	Time (s)
Our Systems	Guanhua	73.25	64.48	4,524
	Yue	71.66	63.06	4,191
	Wu	72.41	63.59	4,578
Mono-phone HMMs	Guanhua	68.77	59.44	351
	Yue	67.54	58.38	336
	Wu	67.23	57.53	333
Tri-phone HMMs	Guanhua	74.34	66.17	23,802
	Yue	72.57	64.79	25,161
	Wu	73.38	64.71	26,871

TABLE VII
EVALUATION OF GUANHUA SYSTEM ON YUE AND WU ACCENT

System	Guanhua		Guanhua + Yue ADM
Testing Set	TestY	TestW	TestY
Phone Correct%	66.22	63.92	68.16
Phone Accuracy%	52.15	51.18	59.21

similar phones in terms of the articulatory events provided by the detectors. Meanwhile, our systems perform a 1.51% absolute phone accuracy reduction but 5.71 times faster than triphone HMM systems on average. Our systems can be further improved by incorporating more acoustics and phonetics-motivated features, which we will study in the future.

B. Accent Related Attribute Discrimination Module Integration

A comparison of Table VI to Table VII shows that Yue and Wu accent variations severely degrade the system trained on Guanhua. In such scenario, we propose integrating an accent-related attribute discrimination module to the system to cover accent changes without retraining any model of the system. Such module aims at reducing the confusion between certain articulatory features that causes phoneme variations.

For a specified confusing articulatory attribute, a support vector machine was trained with samples of such features in the Yue accent. SVM is an ideal binary classifier in terms of minimum structural risk, and has been successfully used in ASR [23]. When testing on the Yue accent data, we use the SVM to reclassify the frames of the detected events that correspond to the confusing articulatory attributes, then replace those events with the reclassification result, and deliver them to CRF as usual. We built a module for palato-alveolar/dental confusion using samples from TrainY. We use the radial basis function kernel in SVM [23]. The inputs to SVM are also $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. The corresponding phone recognition results are listed in Table VII.

With the accent related attribute discrimination module, the accuracy of palato-alveolar and dental increased from 71.0% and 76.7% to 83.0% and 92.1%, correspondingly. The absolute phone accuracy of the Guanhua system was significantly improved by 7.06%, which shows the effectiveness of the proposed attribute discrimination module that is trained with only 6.49% of the samples in the TrainY set (118,751 among 1,830,778 samples). Such module can be used for "plug 'n' play" since we can employ or not employ a module for any accent and attribute in a system without retraining any model, which enhances the system flexibility and is easy for studying specific accent changes.

By analyzing the phone confusion matrix, we found a batch of accent changes related to palato-alveolar/dental confusion have been reduced, for example, 'zh' to 'z', 'ch' to 's' and 'sh' to 's', which change the most in accent variations and are the most difficult to be correctly recognized in the Yue accent found by acoustic phoneticians [1]. Meanwhile, the correctness of some vowels not directly related to the palato-alveolar/dental confusion has also been improved by reducing their deletion errors. These results indicate that, according to underlying data, CRF integrates the detected events following but not limited by designated linguistic knowledge. Improvements of some phones are illustrated in detail in Figure 3.

TABLE VIII
COMPARISIONS OF MULTI-ACCENT ROBUST SYSTEM

System	Accent	Phone Correct%	Phone Accuracy%	Time (s)
Our Multi-accent System	Guanhua	73.83	65.39	13,128
	Yue	73.75	66.49	13,384
	Wu	73.13	67.09	12,740
Triphone HMMs	Guanhua	74.49	65.40	45,208
	Yue	74.79	66.34	46,740
	Wu	73.70	65.87	43,008

C. A Robust Multi-Accent System

Next we presented a method of building a robust multi-accent ASR system. As we have discussed, detectors are influenced by regular accent variations when detecting on a different accent. Intuitively, detectors trained on different accents display different regularities on the same accent, and it is possible to extract patterns for multi-accent changes from the evidences provided by these detectors. Therefore, we used the Guanhua, Yue and Wu detectors simultaneously in one system, and integrate the detected events in parallel by CRF. We use features elaborated in Section III as well as the features shown in the following:

Interactive Feature: such features are combinations of attributes belonging to the same category relevant to different accents, at the current, the previous two and the next two frames.

Consequently, we employed detectors in Table V together with context-independent phone detectors for each accent. A CRF merger with 788,448 features was constructed. We compare our system with a triphone HMM system trained on all data from TrainG, TrainY and TrainW. The baseline triphone system was trained on the same data set and has 1,203 tied-states, with each state having 6 Gaussian components. The comparisons are listed in Table VIII.

It is shown in Table VIII that our system obtains a comparable

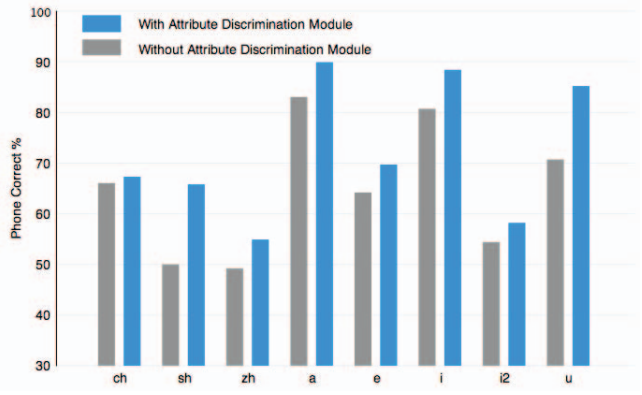

Fig. 3. Comparisons of phone correctness of Guanhua system tested on Yue accent with/without using Yue accent attribute discrimination module

phone accuracy at 3.44 times faster than the triphone HMM system on average, without retraining the detectors.

V. Conclusion

We presented an approach to accented speech recognition using detection based automatic speech attribute transcription (ASAT) with articulatory features. First we used conditional random fields to combine detected events with frame-level input and binary feature function to appropriately utilize articulatory and phonetic knowledge with good scalability. Next we used merely state features in CRF and make recognition decision in the evidence verifier to solve the problem that CRF under-generate phones. We then proposed an accent-related attribute discrimination module to handle a diversity of accent changes as well as to increase the flexibility of the system since no model need to be retrained when using this module. Experimental results showed that our proposed method outperforms the conventional monophone HMM system by 5.04%, 4.68% and 6.06% absolute phone accuracy improvement on Guanhua, Yue and Wu, respectively. Compared to the triphone HMM systems, our approach achieves a comparable phone accuracy at speed of 5.71 times faster. Furthermore, we demonstrated our system is able to cover flexible multi-accent variations in a single system.

Acknowledgment

The first author would like to thank Xuan Wang at Beijing Language and Culture University, advising the linguistic knowledge utilized and helping to improve the English writing. This work was support by Natural Science Foundation of China (60975018), State Scholarship Fund of China Scholarship Council (2010811231), and the joint research grant of Nokia-Tsinghua Joint Funding 2008-2010.

References

[1] J.-H. Yuan *et al*, *Survey of the Chinese Dialects*, Yu Wen Press, Beijing, second edition, 2001, (in Chinese).

[2] Y. Liu and P. Fung, "Partial change accent models for accented mandarin speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, Nov./Dec. 2003, pp. 111–116.

[3] Leading Group Office of Survey of Language User in China, *Survey of Language Use in China*, Yu Wen Press, Beijing, 2006, (in Chinese).

[4] G.-H. Ding, "Phonetic confusion analysis and robust phone set generation for Shanghai-accent Mandarin speech recognition," in *Interspeech-2008*, Brisbane, Australia, Sept. 22-26, 2008, pp. 1129–1132.

[5] Y. R. Oh and H. K. Kim, "MLLR/MAP adaptation using pronunciation variation for non-native speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, Dec. 13–17, 2009, pp. 216–221.

[6] M. Saraclar *et al*, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech Language*, vol. 14, pp. 137–160, Apr. 2000.

[7] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[8] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech Language*, vol. 9, pp. 171–185, Apr. 1995.

[9] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Interspeech-2004*, Jeju Island, Korea, Oct. 4–8, 2004.

[10] C.-H. Lee *et al*, "An overview on automatic speech attribute transcription (ASAT)," in *Interspeech-2007*, Antwerp, Belgium, Aug. 27–31, 2007, pp. 1825–1828.

[11] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in *Interspeech-2006*, Pittsburgh PA, USA, Sept. 17–21, 2006, pp. 1287–Mon3BuP.3.

[12] S. M. Siniscalchi *et al*, "Toward a detector-based universal phone recognizer," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, Mar./Apr. 2008, pp. 4261–4264.

[13] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, USA, Dec. 12-15, 1999, pp. 79–84.

[14] E. C. Sagey, *The Representation of Features and Relations in Non-linear Phonology*, Ph.D. thesis, Dept. of Linguistics and Philosophy, MIT, Cambridge, 1986.

[15] C.-Y. Lin and H.-C. Wang, "Attribute-based Mandarin speech recognition using conditional random fields," in *Interspeech-2007*, Antwerp, Belgium, Aug. 27-31, 2007, pp. 1833–1836.

[16] J. Lafferty *et al*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning*, Williamstown, MA, USA, June 28-30, 2001, pp. 282–289.

[17] C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning*, MIT Press, Cambridge, 2006.

[18] ChineseLDC.org, *Introduction to RASC863*, http://www.chineseldc.org/doc/CLDC-SPC-2004-005/intro.htm, 2009.

[19] C. Zhang *et al*, "Reliable accent specific unit generation with dynamic Gaussian mixture selection for multi-accent speech recognition," in *Proc. IEEE International Conference on Multimedia & Expo*, Barcelona, Spain, July 11-15, 2011.

[20] S. Young *et al*, *The HTK Book*, Cambridge Research Laboratory, Cambridge, UK, 3.4 edition, 2009.

[21] T. Kudo, *CRF++: Yet Another CRF Toolkit*, http://crfpp.sourceforge.net, 2010.

[22] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, Apr. 2011.

[23] O. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, USA, Mar. 15-19, 1999, pp. 585–588.