Integration of Articulatory Knowledge and Voicing Features Based on DNN/HMM for Mandarin Speech Recognition

Ying-Wei Tan, Wen-Ju Liu, Wei Jiang and Hao Zheng

Abstract— **Speech production knowledge has been used to enhance the phonetic representation and the performance of automatic speech recognition (ASR) systems successfully. Representations of speech production make simple explanations for many phenomena observed in speech. These phenomena can not be easily analyzed from either acoustic signal or phonetic transcription alone. One of the most important aspects of speech production knowledge is the use of articulatory knowledge, which describes the smooth and continuous movements in the vocal tract. In this paper, we present a new articulatory model to provide available information for rescoring the speech recognition lattice hypothesis. The articulatory model consists of a feature front-end, which computes a voicing feature based on a spectral harmonics correlation (SHC) function, and a back-end based on the combination of deep neural networks (DNNs) and hidden Markov models (HMMs). The voicing features are incorporated with standard Mel frequency cepstral coefficients (MFCCs) using heteroscedastic linear discriminant analysis (HLDA) to compensate the speech recognition accuracy rates. Moreover, the advantages of two different models are taken into account by the algorithm, which retains deep learning properties of DNNs, while modeling the articulatory context powerfully through HMMs. Mandarin speech recognition experiments show the proposed method achieves significant improvements in speech recognition performance over the system using MFCCs alone.**

I. Introduction

In spontaneous speech, there is much variability that makes a significant challenge to the performance of state-of-the-art continuous automatic speech recognition (ASR) systems [1]. Such variability mainly origins from coarticulation [2] and it has been proposed that coarticulation can be effectively coped with by applying articulatory knowledge [3], [4]. In [5], [6], articulatory based processing algorithms have been proposed to model the acoustic-phonetic variations. In [7], articulatory features are hypothesized to capture acoustic variation at finer level than the phoneme-based representation. In [8], representations of speech production are used to improve automatic speech recognition (ASR).

Speech modelling generally occurs in the acoustic domain. A speech recognition system must take acoustic signals as input, however to take these without considering speech production mechanism ignores a rich source of prior knowledge [9]. Several studies have indicated that articulatory input representation contains available information which is complementary to that provided by standard MFCC features [10]. In [11], techniques based on acoustic-to-articulatory feature codebooks are applied to accurate recovery of articulator positions. In [12], a phone-based background model (PBM) approach is presented to improve attribute detection accuracies. Methods based on artificial neural networks (ANNs) are exploited to extract articulatory features from the acoustic speech signal [13]. In the algorithm, articulatory features can be estimated by multi-task learning (MTL) multilayer perceptrons (MLPs) compactly and efficiently, and the inter-feature dependencies are learned through a common hidden layer representation. Furthermore, adding phoneme as subtask while estimating articulatory features improves both articulatory feature estimation and phoneme recognition. In [14], a probabilistic framework for landmark-based speech recognition that utilizes the sufficiency and context invariance properties of acoustic cues for phonetic features is presented. In [15], support vector machines (SVMs) are explored to capture fine phonetic variation in speech using articulatory features. In [16], a flexible stream architecture based on Gaussian mixture models (GMMs) is used for automatic speech recognition (ASR) with articulatory features. In [17], the method determines articulatory movements from speech acoustics using a hidden Markov model (HMM) based speech production model. The model statistically generates speech spectrum and articulatory parameters from a given phonemic string. In [18], neural networks are trained to map short-term spectral features to the posterior probability of some distinctive features. These probabilities are then used as features in a large vocabulary tied-state HMM-based recognition. Considering factored models of the articulatory state space with an explicit model of articulator asynchrony, a factored conditional random fields (CRFs) are applied for articulatory feature forced transcription [19]. In [20], [21], dynamic Bayesian networks (DBNs) model articulatory-acoustic context with an auxiliary variable that complements the phonetic state variable. In [22], dynamic Bayesian networks (DBNs) are used to run articulatory feature recognition in conjunction with an embedded training scheme designed to learn asynchronous feature value changes. In [23], applying articulatory features obtained by hierarchical multilayer perceptron (MLP) improves the discrimination of tone modeling effectively. In [1], a deep neural network (DNN) is exploited to extract articulatory information from the speech signal in a continuous speech recognition task and demonstrated that with deeper networks the performance of speech recognition systems can be improved.

As mentioned above, these methods only take advantages of a single model. To fully exploiting the virtues of various

YingWei Tan, WenJu Liu, Wei Jiang and Hao Zheng are with the Department of National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (email: {ywtan,lwj,wjiang,hao.zheng}@nlpr.ia.ac.cn).

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267, No.90820011 and No.90820303).

978-1-4799-1959-8/15/$31.00 ©2015 IEEE

models, the fusing of different models is received with concern. In [24], algorithms based on discriminative model combination (DMC) are used to integrating multilingual articulatory features into speech recognition. The algorithm enables us to better address the problem of non-native speech recognition. In [25], the frame-level classification of a set of articulatory features (AFs) inspired by the vocal tract variables of articulatory phonology is studied. The algorithm performs the incorporation of k nearest neighbors and multilayer perceptrons (MLPs) for articulatory feature classification. In [26], the combination of artificial neural networks and dynamic Bayesian networks is made to perform articulatory feature recognition. In [27], a hybrid hidden Markov models (HMMs)/Bayesian networks (BNs) model is proposed to effectively utilize the available articulatory information for improving the performance of automatic speech recognition (ASR).

For tonal languages such as Mandarin, it is widely known that tone and voicing information can help to improve the automatic speech recognition (ASR) performance [28]. Voicing feature extraction is a key method in finding discriminative cues for Mandarin speech recognition. Previous work in incorporating voicing features into speech recognition systems includes the following. In [29] methods with autocorrelation based voicing features are presented in a HMM-based speech recognition system. In [30] pitch and a voicing feature are combined with standard MFCCs using linear discriminate analysis (LDA). In [31] three alternative voicing features are reported in combination with MFCCs features using the LDA algorithm. They all show improvements in speech recognition accuracy rates. In [32] the entropy of the high-order cepstrum and the normalized autocorrelation peak are combined with MFCCs using HLDA [33] to obtain discriminative cues. In [34] voicing features and spectrum derivative features are combined with MFCCs using LDA. The method also achieves some improvements in speech recognition accuracy rates.

In this work, we propose a method which formulates a new voicing feature extraction algorithm using the spectral harmonics correlation function in frequency domain and build a articulatory model using hybrid deep neural networks (DNNs) and hidden Markov models (HMMs) methods to map both the acoustic and voicing features into articulatory space. The method provides great discriminative cues for Mandarin speech recognition. As for the voicing feature extraction algorithm, firstly, nonlinear preprocessing is performed for restoring the fundamental and enhancing the voicedness in high frequency domain. Secondly, we obtain a new voicing feature through arranging the formulation of the SHC function. Then, the voicing feature is normalized by cepstral variance normalization (CVN) to reduce the residual mismatch in each utterance. Finally, the normalized feature and its derivatives are incorporated with MFCCs using HLDA to obtain the most relevant features. As will be shown in the paper, we evaluate the resulting features on a Mandarin speech recognition task. The experimental results show the algorithm achieves significant CER reductions for the task. It is relatively up to 20.73%. In addition, the resulting features captured by HLDA are used to train hybrid DNN/HMM models, which are used to incorporate articulatory knowledge into speech recognition systems by rescoring the speech recognition lattice hypotheses. Observation posterior probabilities, initial state probabilities and state transition probabilities of articulatory models are incorporated into the speech recognition lattice rescoring process. It is observed that applying articulatory knowledge based on hybrid DNN/HMM models can improve speech recognition rates on the Mandarin speech recognition tasks much further. The best result is reported with a 22.75% relative reduction of CER.

The outline of this work is as follows. In Section 2 we describe the formulation of voicing feature extraction and the combination of voicing features and MFCCs using the HLDA algorithm in detail. In the following Section we show the articulatory modeling process using a hybrid DNN/HMM framework and the lattice rescoring formulation based on articulatory models. In Section 4 we show the speech recognition task setups, the speech databases and experimental results on evaluating our algorithm. Finally the conclusions are drawn.

II. Feature description

In this Section, firstly, we introduce how new voicing features based on the SHC function are formulated in detail. Secondly, we describe the combination of the proposed voicing features and MFCCs using the HLDA algorithm.

A. Voicing feature extraction

Common inspiration of voicing feature extraction methods is to detect the quasi periodic oscillation of the vocal chords and capture discriminative cues between voiced and unvoiced sounds. Spectral harmonics correlation (SHC) functions based methods measure the periodicity of a time frame in the frequency domain. The specific steps are described as follows:

First, nonlinear preprocessing is performed to compensate the calculation of the SHC function by using the squared value. The methods can be used to partially restore a missing fundamental. The restoration of the fundamental by using the squaring operation is also illustrated by using spectrograms in Figure 1. The top panel (a) depicts the spectrogram of a speech signal, for which the fundamental above 500 Hz is not clearly apparent enough. In contrast, the fundamental is more clearly apparent in the spectrogram of the nonlinearly processed signal shown in the bottom panel (b).

Then, the nonlinearly processed speech signal is bandpass-filtered for reducing the magnitude of the DC component. The bandwidths (50-1500 Hz) and orders (150 points) of the bandpass finite impulse response (FIR) filters are determined empirically.

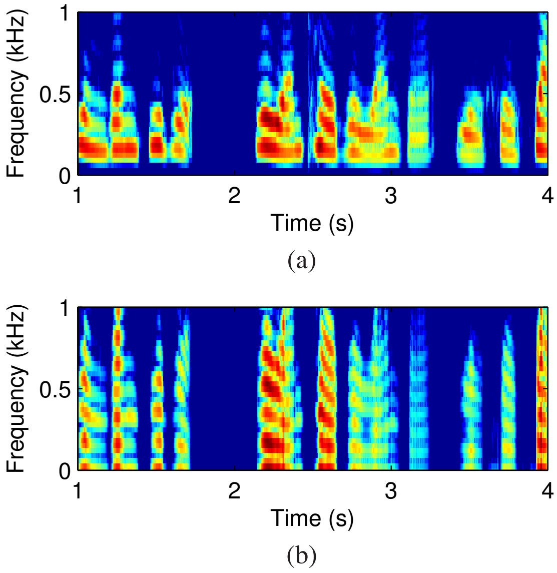


Fig. 1. The spectrogram of a clean Mandarin speech signal, (a) before applying nonlinear processing and (b) after applying nonlinear processing.

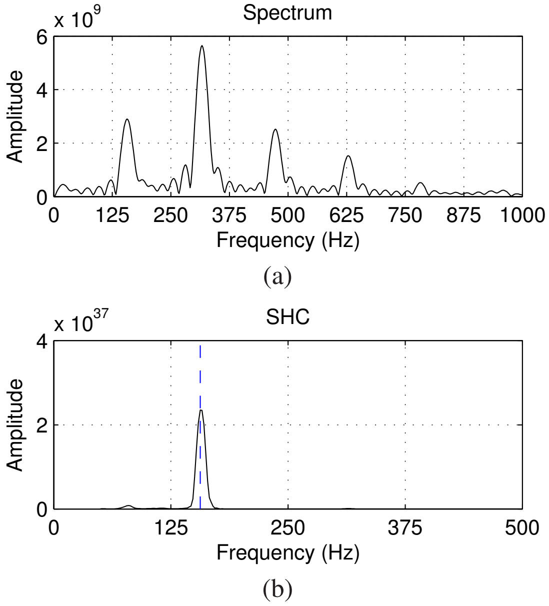


Fig. 2. (a) The spectrum and (b) the corresponding spectral harmonics correlation of a voiced speech frame.

Next, the spectral harmonics correlation (SHC) function is defined to use multiple harmonics in [35] as follows:

$$\mathrm{shc}(t,f) = \sum_{f'=-n/2}^{n/2} \prod_{i=1}^{h+1} Y(t, i\cdot f + f'), \qquad (1)$$

where $Y(t,f)$ is the magnitude spectrum for frame t at frequency f, n is the spectral window length in frequency, and h is the number of harmonics. f is a discrete variable with a spacing dependent on fast Fourier transformation (FFT) length N_f and sampling rate f_s. For each frequency f, $Y(t,f)$ represents the extent to which the spectrum has high amplitude at integer multiples of that f. Empirically, $n = 40$ Hz and $h = 3$. $Y(t,f)$ results in prominent peaks at the fundamental frequency.

Figure 2 shows the spectrum and the corresponding spectral harmonics correlation of a voiced speech frame. Compared to the small peak at the fundamental frequency of around 156 Hz in the spectrum, a very prominent peak is observed in the spectral harmonics correlation function.

The aim of voicing feature extraction is to produce a bounded value describing how voiced the current frame is, we develop a new voicing feature that evaluates the peak structure of SHC. Voiced frames exhibit a sharp maxima, while unvoiced frames have no clear peak structure. The feature v_{shc} evaluates the maximum amplitude value of SHC. It is defined as the ratio of the maximum amplitude value and the algebraic mean of the neighboring amplitudes without the maximum value.

$$f_{\max} = \mathop{\arg\max}_{\mathrm{F0\_min} \le f \le \mathrm{F0\_max}} \mathrm{shc}(t,f), \qquad (2)$$

$$v_{\mathrm{shc}} = \frac{\mathrm{shc}(t, f_{\max})}{\sum_f \mathrm{shc}(t,f) \Big/ 2W}. \qquad (3)$$

Because an adult's pitch frequency commonly ranges from 50 Hz to 400 Hz, the calculation is performed only for the designated search range $\mathrm{F0\_min} \le f \le \mathrm{F0\_max}$, with $\mathrm{F0\_min} = 50$ Hz and $\mathrm{F0\_max} = 400$ Hz. The algebraic mean is calculated over the neighborhood of $f_{\max}$. The size of the neighborhood is set to $W = \left\lfloor \mathrm{F0\_min} \Big/ \frac{f_s}{N_f} \right\rfloor$ to avoid peaks of the neighboring harmonics being included in the average. f runs from $f_{\max} - W \cdot \frac{f_s}{N_f}$ to $f_{\max} + W \cdot \frac{f_s}{N_f}$ excluding $f_{\max}$. Typically we have $1 \le v_{\mathrm{shc}} \le 12$.

Figure 3 depicts distributions of v_{shc} on voiced and unvoiced sounds. The histogram of a given phoneme has been estimated on values aligned to the given phoneme on the "863" train set. The distributions of v_{shc} reveals the distinction of voiced and unvoiced sounds.

B. The combination of voicing features and MFCCs

Linear discriminant analysis (LDA) is broadly applied to reduce dimensionality and a powerful method to preserve discriminative information. LDA assumes each class has the same class covariance. However, this assumption does not necessarily hold for a real data set. In order to remove this limitation, heteroscedastic linear discriminant analysis (HLDA) has been presented. Heteroscedastic linear discriminant analysis (HLDA) can deal with unequal class covariances because the maximum likelihood estimation is used to estimate parameters for different Gaussians with unequal class covariances. Here, the combination of voicing features and MFCCs using the HLDA algorithm are described as follows:

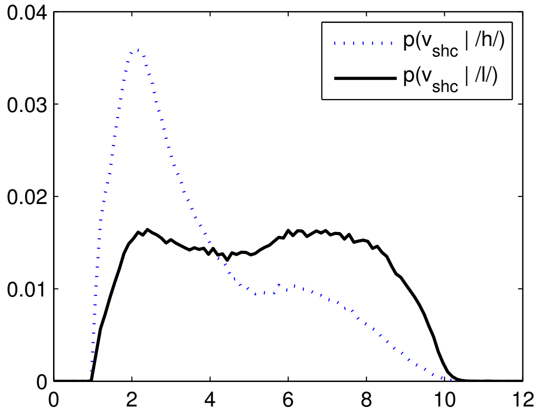


Fig. 3. Histograms of the feature v_{shc} estimated on “863” corpus. The solid line corresponds to v_{shc} of voiced initial ‘l’ and the dot line corresponds to the unvoiced initial ‘h’.

Firstly, the extracted voicing feature v_{shc} is normalized as $\hat{v}_{\text{shc}}$ in each utterance so that each value has zero mean and unit variance. The normalized feature $\hat{v}_{\text{shc}}$ and its time derivatives are added into the standard acoustic feature vector (MFCCs). Then, HLDA is used to project 42 dimension to 39 dimension with reserving the most relevant classification information. Finally, the resulting 39 dimension features are used as the input features of the subsequent articulatory modeling.

III. Articulatory modeling using DNN/HMM

In this Section, firstly, we introduce deep neural networks briefly. Secondly, we describe hidden Markov models shortly. Next, articulatory modeling are elaborated using the hybrid DNN/HMM framework. Finally, the lattice rescoring process based on articulatory models is described.

A. Deep neural networks

The DNN framework has become the dominant techniques in acoustic modelling in speech recognition [36]. This acoustic modeling technique differs from the earlier ANN systems in that there are more hidden layers and more hidden nodes in each hidden layer in the DNN topology.

Given an input observation vector $\mathbf{o}$, DNNs pass it through multiple hidden layers $\mathbf{h}_i$, where $i = 1, \cdots, L$ and L is the number of hidden layers. This procedure can be formulated as follows:

$$\mathbf{h}_1 = \sigma(\mathbf{W}_1 * \mathbf{o} + \mathbf{b}_1), \tag{4}$$

$$\mathbf{h}_{i+1} = \sigma(\mathbf{W}_i * \mathbf{h}_i + \mathbf{b}_i),\ i = 1,\ \cdots,\ L, \tag{5}$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight matrix and bias vector for the hidden layer respectively. σ is the sigmoid function. For the output layer, the softmax function

$$p(y = a|\mathbf{h}_L) = \frac{\exp(\mathbf{W}_L^a * \mathbf{h}_L + \mathbf{b}_L^a)}{\sum\limits_{y'} \exp(\mathbf{W}_L^{y'} * \mathbf{h}_L + \mathbf{b}_L^{y'})}, \tag{6}$$

is used to estimate the label posterior probability $p(y = a|\mathbf{o})$ where a represents the articulatory class label, $\mathbf{W}_L^{y'}$ and $\mathbf{b}_L^{y'}$ are the y'th row of the weight matrix $\mathbf{W}_L$ and the y'th element of bias vector $\mathbf{b}_L$.

B. Hidden Markov models

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with hidden states. A HMM, defined as $\lambda = \{A, B, \pi\}$, consists of the following elements:

- The number of states in the model denoted as N, the set of states denoted as $S = \{s_1, s_2, \cdots, s_N\}$ and q_t the state at time t.
- $A = \{\mathrm{a}_{ij}\}$, the state transition probability with

$$\mathrm{a}_{ij} = P(q_t = s_j|q_{t-1} = s_i), 1{\le}i, j{\le}N. \tag{7}$$

- $B = \{b_i(\mathbf{o}_t)\}$, the observation probabilities, where $b_i(\mathbf{o}_t)$ represents the probability of observation $\mathbf{o}_t$ at state s_i.
- $\pi = \{\pi_i\}$, the initial state probabilities, where $\pi_i = P(q_1 = s_i), 1{\le}i{\le}N$.

To make the HMM available, there are two problems that one should solve: [37]

- Learning problem: Given the observation sequence $\mathbf{O} = \mathbf{o}_1\mathbf{o}_2\cdots\mathbf{o}_T$, the learning procedure is to find the set of model parameters $\lambda^* = \{A^*, B^*, \pi^*\}$, such that $\lambda^* = \arg\max_\lambda P(\mathbf{O}|\lambda)$. The Baum-Welch algorithm is employed to solve the learning problem.
- Decoding problem: Given a model λ and a sequence of new observations $\mathbf{O} = \mathbf{o}_1\mathbf{o}_2\cdots\mathbf{o}_T$, the decoding procedure is defined as the problem of finding the hidden state sequence that have most likely produced that observation $q_1q_2\cdots q_T$. The solution of this problem is given by the Viterbi algorithm [37] as

$$P(\mathbf{O}|\lambda) = \max_{q_1q_2\cdots q_T} \pi_{q_1} \prod_{t=2}^{T} P(q_t|q_{t-1}) b_{q_t}(\mathbf{o}_t). \tag{8}$$

C. DNN/HMM for articulatory modeling

In the case of articulatory modeling, we need to understand articulatory knowledge. For Mandarin speech, Chinese syllable pertains to the Initial-Final structure. Both the initials and finals can be further divided into several detailed categories according to the manner and the place of articulation. The categories are described in Table I. We assume that class labels consists of 19 articulatory categories, a pseudo initial class and a silence class. For each phoneme, the mean of the input features is used for articulatory modeling in its corresponding duration. In the training stage, the training speech data is aligned and its corresponding labels is used to train a universal DNN whose output layer has N output nodes where $N = 21$ is the number of class labels. In the testing stage, the testing speech data is passed to the resulting DNN to compute the posterior probability of each phoneme. After calculating the posterior probabilities for a phoneme, the label of the phoneme is determined as the articulatory label which has the maximum posterior probability.

TABLE I

Articulatory categories of initials and finals.

	Categories	Description	
1	*m n l r y w*	Voiced	Initial
2	*b p d t g k*	Stop	
3	*z c zh ch j q*	Fricative	
4	*f s sh x h*	Affricate	
5	*a ia ua*	Simple vowel and tail-dominant	Final
6	*e ie üe*		
7	*o uo*		
8	*i*		
9	*u*		
10	*ü*		
11	*er*		
12	*ai uai*	head-dominant and centre-dominant	
13	*ei uei*		
14	*ao iao*		
15	*ou iou*		
16	*an ian üan uan*	Nasal	
17	*in en uen üen*		
18	*ang iang uang*		
19	*eng ong ing iong*		

Taking the context of articulatory information into account, we build a HMM model with 21 states symbolizing articulatory categories. The HMM model is ergodic. The key point in the DNN/HMM framework is using the DNN's posterior probability to represent the observation probabilities $b_i(\mathbf{o}_t)$. According to [37], the estimation formulas of the initial state probabilities and the state transition probability can be derived as

$$\bar{\pi}_i = \gamma_1(i), \tag{9}$$

$$\bar{\mathrm{a}}_{ij} = \frac{\sum\limits_{t=1}^{T-1} \xi_t(i,j)}{\sum\limits_{t=1}^{T-1} \gamma_t(i)}, \tag{10}$$

where $\xi_t(i,j)$ is the probability of being in state s_i at time t and state s_j at time $t-1$ and $\gamma_t(i)$ is the probability of being in state s_i at time t, given the model λ and the observation sequence $\mathbf{O}$.

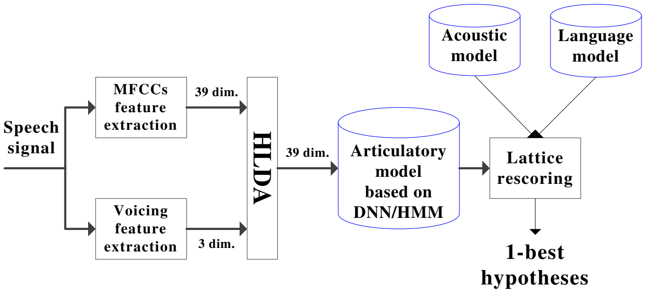


Fig. 4. The lattice rescoring process of articulatory models based on the hybrid DNN/HMM framework.

D. Integrating articulatory models into the speech recognition

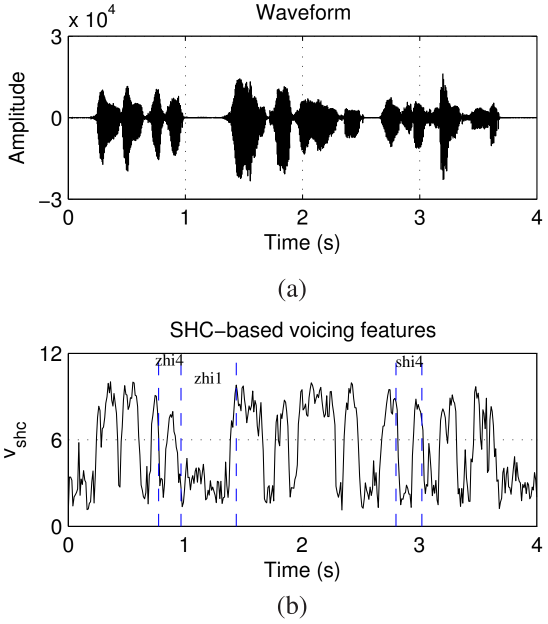
Figure 4 shows the flow diagram of the lattice rescoring process of articulatory models based on hybrid DNN/HMM models. The articulatory models are integrated into the continuous speech recognition system by rescoring the lattice hypotheses. The lattice hypotheses are generated by the first pass recognition, and are reranked with the articulatory scores integrated. The total score of a hypothesis can be defined as:

$$\Phi = \sum_{i=1}^{I} [\Phi_{\mathbf{AM}}(p_i) + \alpha\Phi_{\mathbf{LM}}(p_i) + \beta\Phi_{\mathbf{DNN}}(p_i) + \delta\Phi_{\mathbf{HMM}}(p_i)], \tag{11}$$

where α is the language model weight, β is the weight correspond to the DNN's posterior probability of articulatory models, δ is the weight corresponding to the HMM's initial state probability or the HMM's state transition probability of articulatory models, and I is the phoneme number of the hypothesis for a candidate path. Thus, the hypothesis with the highest path score is regarded as the best hypothesis.

IV. Experiments and Results

The details of generation of the new voicing feature v_{shc} are summarized in this section. For every 10 ms, a 40 ms long window is applied to the speech signal. The window is longer than for MFCCs to increase the possible number of periods in a time frame. Before computing the SHC, nonlinear preprocessing is performed and Kaiser window is used. To increase the frequency resolution, an 8192-point FFT is computed with zero padding. The SHC-based voicing features for a waveform extracted from the Mandarin speech corpus are shown as Figure 5, where the curves of v_{shc} reveal the discrimination of voiced and unvoiced sounds. In addition, it is observed that different syllables with the same tone have different v_{shc} contour and different tones with the same syllable make v_{shc} contour different in shape.



(a)

(b)

Fig. 5. (a) The waveform and (b) the corresponding SHC-based voicing features for a clean Mandarin speech signal.

Experiments are performed on the “863” corpus, which is provided by Chinese National Hi-Tech Project “863” for Mandarin large vocabulary continuous speech recognition. 83 male speakers’ data is employed for training (48373 sentences, 55.6 hours) and 6 male speakers’ for test (240 sentences, 17.1 minutes). For “863” speech data, the acoustic models are trained as triphone HMMs with decision tree-based state clustering and each state is modeled by a 16-component Gaussian mixture model. The model uses three states (left-to-right) per phone. The system uses a bigram language model with 48188 words. Training and decoding are performed using HTK tools [38].

In the baseline system, acoustic features (MFCCs) are 12 dimension MFCC plus 1 normalized energy and their 1st and 2nd order derivatives and cepstral mean normalization (CM-N) is applied. Acoustic models are trained with maximum likelihood estimation (MLE).

In the following subsection, experimental results are given to illustrate the performance of the proposed voicing features. Besides, the voicing features are combined with tone features, which are similar to those described in [39], including the general spline interpolation, moving window normalization and 5-point moving average smoothing.

TABLE II

CER (%) ON THE “863” SPEECH CORPUS.

Acoustic features	Dim.	CER
MFCCs	39	12.88
MFCC+v_{hps}	40	12.46
MFCCs+v_{shc}	40	12.43
MFCCs+$\hat{v}_{\text{shc}}$	40	11.63
MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)	42	11.07
MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA	39	**10.21**
MFCCs+(F0,ΔF0,$\Delta\Delta$F0)+HLDA	39	13.16
MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+(F0,$\Delta$F0,$\Delta\Delta$F0)+HLDA	39	12.85

A. *speech recognition using voicing features*

The results of the experiments for the systems are given in Table II. Firstly, it can be easily observed that the single feature v_{shc} is catenated with the standard MFCCs to improve recognition rates directly. In frequency domain, the SHC-based methods give better results over the HPS-based methods [31]. After v_{shc} is normalized to reduce the residual mismatch, recognition rates rise much further. Adding time derivatives of the normalized feature is also useful in recognition rates. Then, the HLDA transform is used to project 42 dimension to 39 dimension with reserving the most relevant classification information. The resulting feature vector has same size to ensure comparable recognition results. In this case, we make continuous progress and obtain a relative implements in character error rate (CER) of 20.73%. Besides, in contrast with tone features, our algorithm show greater advantages in recognition rates. The possible reason is that SHC-based features are continuous naturally while tone-based features are interpolated by force at unvoiced regions. Finally, HLDA makes the combination of voicing features and tone features and projects 45 dimension to 39 dimension. However, no improvements are found in this case.

B. *Articulatory recognition*

Using the baseline recognition system, we perform the force alignment, which is applied to train the articulatory model. For the DNN training, the training data is randomly divided into the train and validation set. The articulatory recognizers are tested on 240 utterances from the test set. The statistics of data used in the experiments is listed in Table III. For the DNN-based system, we examine different configurations of the DNNs. The number of hidden layers varies from 3 to 4 and the number of nodes in each hidden layer increases from 512 to 1024. The input layer has 39 visible units and the output layer has 21 nodes. The networks are trained by the method proposed in [40], [41]. DNNs are trained using mini-batch stochastic gradient descent with the batch size being 1000. During the discriminative pretraining, the initial learning rate is set to 0.5. Momentum is used to speed up learning. The momentum starts off at 0.5 and increases linearly to 0.97 over the 50 epochs.

TABLE III

STATISTICS OF DATA USED IN OUR EXPERIMENTS.

Data set	No of samples
train	1266458
dev	6165
test	6165

The experimental results based on DNN are shown in Table IV. It can be seen that the accuracies are improved when the proposed voicing features are fused with traditional MFCCs. When the DNN models have 4 hidden layers which has 1024 nodes in each hidden layer, the classifier achieves the best performance. Thus, we choose the configure for the lattice rescoring.

TABLE IV

ARTICULATORY RECOGNITION RATES (%) OF DNN.

Features	No of layers	No of hidden layer nodes	
		512	1024
MFCCs	3	87.98	88.71
	4	88.94	89.16
MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA	3	88.71	89.27
	4	88.86	**89.29**

Besides, we perform the articulatory classification based multilayer perceptrons (MLPs) that is shallow models. The result is shown in Table V. Also, It can be seen that the accuracies are improved when the proposed voicing features are combined with traditional MFCCs. Compared to Table IV, the performance of the DNN-based method has significant improvement in comparison with the MLP-based algorithm.

TABLE V

ARTICULATORY RECOGNITION RATES (%) OF MLP.

Features	Results (%)
MFCCs	86.49
MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA	**86.83**

C. Rescoring the lattice hypotheses

As can be seen in Table VI, when using MFCCs, the DNN-based method is better than the MLP-based method. However, the improvement is not obvious. When using the proposed features, the performance of the DNN-based method is equivalent to that of the MLP-based method. Compared to the baseline system, when the articulatory model based on the hybrid DNN/HMM framework is merged with the weights ($\beta = 2$ and $\delta = 2$), a 6.91% relative reduction is gained. When voicing features are incorporated into systems also, the best results are achieved with a 22.75% relative reduction. Besides, we can see that the role of the DNN's posteriori probability $\Phi_{\mathbf{DNN}}$ is greater than that of the HMM's state transition probability $\Phi_{\mathbf{HMM}}$ in the experiments. The base 10 logarithm of these probability value is taken.

TABLE VI

CER (%) WITH ARTICULATORY MODELS.

Systems	**CER**
MFCCs + $\Phi_{\mathbf{MLP}}$ (with MFCCs only)	12.15
MFCCs + $\Phi_{\mathbf{DNN}}$ (with MFCCs only)	12.08
MFCCs + $\Phi_{\mathbf{HMM}}$ (with MFCCs only)	12.59
MFCCs + ($\Phi_{\mathbf{DNN}}$ + $\Phi_{\mathbf{HMM}}$) (with MFCCs only)	11.99
(MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA) + $\Phi_{\mathbf{MLP}}$ (with (MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA))	10.05
(MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA) + $\Phi_{\mathbf{DNN}}$ (with (MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA))	10.05
(MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA) + $\Phi_{\mathbf{HMM}}$ (with (MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA))	10.17
(MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA) + ($\Phi_{\mathbf{DNN}}$ + $\Phi_{\mathbf{HMM}}$) (with (MFCCs+($\hat{v}_{\text{shc}}$,$\Delta\hat{v}_{\text{shc}}$,$\Delta\Delta\hat{v}_{\text{shc}}$)+HLDA))	**9.95**

V. CONCLUSIONS

In this work, the integration of articulatory knowledge and voicing features based on hybrid DNN/HMM architectures is presented for Mandarin speech recognition. In this method, a SHC-based normalized feature and its time derivatives are combined with standard MFCCs using HLDA and the hybrid DNN/HMM models with articulatory knowledge are built using the proposed feature set. Experiments performed on large vocabulary Mandarin speech recognition tasks achieve a 22.75% relative reduction of CER. The results demonstrate that the combination of voicing features and articulatory models help in improving Mandarin speech recognition performance.

REFERENCES

[1] V. e. a. Mitra, "Articulatory features from deep neural networks and their role in speech recognition," in *Proceedings of ICASSP*, 2014.

[2] R. Daniloff and R. Hammarberg, "On defining coarticulation," *Journal of Phonetics*, vol. 1, no. 3, pp. 239–248, 1973.

[3] K. N. Stevens, "Toward a model for speech recognition," *The Journal of the Acoustical Society of America*, vol. 32, no. 1, pp. 47–55, 1960.

[4] K. e. a. Livescu, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 jhu summer workshop," in *Proceedings of ICASSP*, vol. 4. IEEE, 2007, pp. IV–621.

[5] K. Kirchhoff, "Robust speech recognition using articulatory information," *PhD Thesis*, 1999.

[6] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2702–2719, 1994.

[7] M. e. a. Richardson, "Hidden-articulatory markov models for speech recognition," *Speech Communication*, vol. 41, no. 2, pp. 511–529, 2003.

[8] S. e. a. King, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[9] J. Frankel and S. King, "Asr-articulatory speech recognition," in *Proceedings of EUROSPEECH*, 2001.

[10] K. e. a. Kirchhoff, "Conversational speech recognition using acoustic and articulatory input," in *Proceedings of ICASSP*, vol. 3. IEEE, 2000, pp. 1435–1438.

[11] J. e. a. Hogden, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1819–1834, 1996.

[12] S. M. e. a. Siniscalchi, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 875–887, 2012.

[13] R. Rasipuram and M. Magimai-Doss, "Improving articulatory feature and phoneme recognition using multitask learning," in *Artificial Neural Networks and Machine Learning*. Springer, 2011, pp. 299–306.

[14] A. Juneja and C. Espy-Wilson, "Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition," *From sound to sense*, vol. 50, pp. 151–156, 2004.

[15] O. e. a. Scharenborg, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication*, vol. 49, no. 10, pp. 811–826, 2007.

[16] F. Metze and A. Waibel, "A flexible stream architecture for asr using articulatory features," in *Proceedings of INTERSPEECH*, 2002.

[17] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[18] B. e. a. Launay, "Towards knowledge-based features for hmm based large vocabulary automatic speech recognition," in *Proceedings of ICASSP*, vol. 1. IEEE, 2002, pp. I–817.

[19] R. e. a. Prabhavalkar, "A factored conditional random field model for articulatory feature forced transcription," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 77–82.

[20] G. Zweig and S. Russell, "Speech recognition with dynamic bayesian networks," in *Proceedings of AAAI*, 1998, pp. 173–180.

[21] G. Zweig and S. J. Russell, "Probabilistic modeling with bayesian networks for automatic speech recognition," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, vol. 98, 1998, pp. 3011–3014.

[22] J. e. a. Frankel, "Articulatory feature recognition using dynamic bayesian networks," *Computer Speech & Language*, vol. 21, no. 4, pp. 620–640, 2007.

[23] H. C. et al., "Improved tone modeling by exploiting articulatory features for mandarin speech recognition," in *Proceedings of ICASSP*, 2012, pp. 4741–4744.

[24] S. e. a. Stüker, "Integrating multilingual articulatory features into speech recognition," in *Proceedings of INTERSPEECH*, 2003.

[25] A. B. e. a. Næss, "Articulatory feature classification using nearest neighbors," in *Proceedings of INTERSPEECH*, 2011, pp. 2301–2304.

[26] J. Frankel and S. King, "A hybrid ann/dbn approach to articulatory feature recognition," in *Proceedings of INTERSPEECH*, 2005.

[27] K. e. a. Markov, "Integration of articulatory and spectrum features based on the hybrid hmm/bn modeling framework," *Speech Communication*, vol. 48, no. 2, pp. 161–175, 2006.

[28] H. C.-H. Huang and F. Seide, "Pitch tracking and tone features for mandarin speech recognition," in *Proceedings of ICASSP*, vol. 3, 2000, pp. 1523–1526.

[29] D. L. Thomson and R. Chengalvarayan, "Use of voicing features in hmm-based speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 197–211, 2002.

[30] A. Ljolje, "Speech recognition using fundamental frequency and voicing in acoustic modeling," in *Proceedings of INTERSPEECH*, 2002.

[31] A. Z. et al., "Extraction methods of voicing feature for robust speech recognition," in *Proceedings of INTERSPEECH*, 2003.

[32] M. G. et al., "Voicing feature integration in sri's decipher lvcsr system," in *Proceedings of ICASSP*, 2004.

[33] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition,” *Speech Communication*, vol. 26, no. 4, pp. 283–297, 1998.

[34] D. K. et al., “Articulatory motivated acoustic features for speech recognition,” in *Proceedings of INTERSPEECH*, 2005, pp. 1101–1104.

[35] S. A. Zahorian and H. Hu, “A spectral/temporal method for robust fundamental frequency tracking,” *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[36] H. G. et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[37] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[38] S. Y. et al., “The htk book,” *Cambridge University Engineering Department*, 2009.

[39] X. L. et al., “Improved tone modeling for mandarin broadcast news speech recognition,” in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2006.

[40] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[41] F. e. a. Seide, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011, pp. 24–29.