Transformation-based Accented Speech Modeling using Articulatory Attributes for Non-Native Speech Recognition

Han-Ping Shen, Chung-Hsien Wu and Pei-Shan Tsai
Department of Computer Science and Information Engineering,
National Cheng Kung University,Tainan
{hanpinsheen, chunghsienwu, pshan08}@gmail.com

***Abstract*—This paper presents a transformation-based approach to robust modeling of accented speech based on articulatory attributes for non-native speech recognition. Firstly, a two-stage verification method is used to extract speech segments from the speech input with non-native accent. Secondly, acoustic models of accented speech are transformed from normal models using linear transformation functions selected from a decision tree to deal with the problem of data sparseness. Thirdly, a discrimination function is applied to filter out the models with low recognition discriminability. Experimental results show that the inclusion of acoustic models of accented speech can eliminate recognition degradation in ASR due to non-native accents and the final ASR system can outperform the standard ASR system in recognizing non-native speech.**

I. INTRODUCTION

Research in multilingual speech recognition has gained increasing interest in the past years due to globalization. For multilingual speech recognition, accents produced by non-native speakers generally degrade the accuracy of automatic speech recognition (ASR). To overcome the degradation problem caused by accents has become one of the most important topics in non-native speech recognition.

Recently, research has been proposed to deal with the accent problems. Existing methods vary from collecting data in a specific accent for recognizer training or adaptation. It has been shown that use of accent-specific data can effectively improves recognition rate [1,2]. For model adaptation, Maximum Likelihood Linear Regression (MLLR) is generally adopted to adapt individually to the speakers with different accents. Conventional approaches to acoustic adaptation, such as MLLR and Maximum a Posteriori (MAP), are the most widely used [3] [4]. Nevertheless, these methods can only train or adapt the acoustic models to recognize specific accented speech according to the collected training or adaptation data. On the other hand, collecting all types of accented speech is time-consuming and almost impossible. For these reasons, this study proposes a systematic method to generate accented acoustic models for dealing with data sparseness problem of accented speech. Acoustic models of normal speech and generated acoustic models of accented speech are integrated for recognition of normal and accented speech.

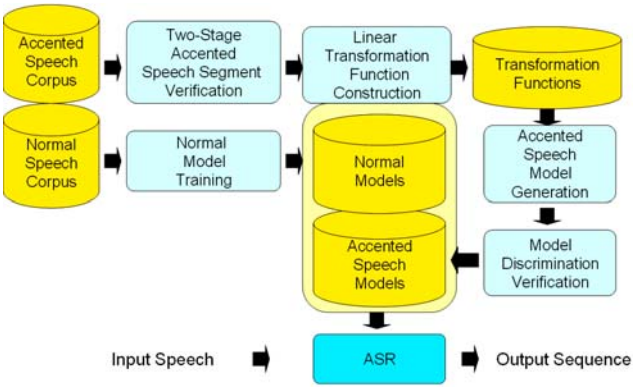

Fig. 1 System Diagram of the Training and Testing Processes.

In this study, a two-stage verification method using acoustic and articulatory attribute-based verification is proposed for acoustic model generation of accented speech. The system diagram of the proposed method is shown in Fig. 1. First, the potential accented speech segments are extracted using the speech recognition confidence and a verifier based on articulatory attribute is proposed for accented speech segment verification. Second, a linear transformation function is constructed based on the pair of a normal pronunciation and the corresponding accented pronunciation. A decision tree is constructed to cluster the accented speech with similar articulatory attributes for transformation function construction to deal with the data sparseness problem and used for conversion of acoustic models of accented speech. Finally, a discrimination function is applied to verify the discriminability of each generated accented speech model so as to filter out the models with low discriminability. The experimental results show the system can obtain acceptable recognition accuracy in recognizing normal speech as well as improve the recognition accuracy in recognizing accented speech.

II. TWO-STAGE ACCENTED SPEECH SEGMENT VERIFICATION

This study proposes a two-stage accented speech segment verification approach to the extraction of accented speech

segments. In the first stage, standard ASR system is adopted for speech recognition. According to the recognition results, speech segments with potential accents can be decided based on the recognition confidence for each recognized phone. Since ASR systems may produce speech recognition error, the second stage is employed to further verify the potential accented speech segments. Some researchers have investigated phonetic variations between accented and non-accented speech [5]. Compared to normal pronunciation, accented speech may lead to different manners or places of articulation. In this study, we assume that accented speech and its corresponding normal speech may have different articulatory attributes (AAs) due to accented pronunciation. An attribute detector [6] is thus used to detect the articulatory attributes of each accented-speech segment. The candidate accented-speech segments are verified to check if the articulatory attributes of the extracted accented speech are different from that of the normal speech.

A. Acoustic feature-based accented speech segment verification

In the first verification stage, an ASR system is adopted to extract potential accented speech segments. Equation (1) and (2) are used to estimate the likelihood of the extracted speech segment x.

$$G_{veri}(x) = \log g(x \mid \lambda_{correct}) - \log g_{Anti-Model}(x) \quad (1)$$

$$g_{Anti-Model}(x) = \frac{1}{N} \sum_{\substack{AAM=1_{\lambda} \\ \lambda_{AAM} \neq \lambda_{correct}}}^{N} g(x \mid \lambda_{AAM}) \quad (2)$$

where $g(.)$ is the recognition likelihood function. $\lambda_{correct}$ is the senone model of the normal pronunciation of speech segment x; a senone is a subphonetic unit denoting a Markov state in phonetic HMMs. λ_{AAM} is the anti-model which is close to $\lambda_{correct}$ and N is the size of the anti-model set. If speech segment x is a normal pronunciation, the value of log $g(x|\lambda_{correct})$ will be high and the value of log $g_{Anti\text{-}Model}(x)$ will be low. This leads to high value of $G_{veri}(x)$. On the other hand, if x is an accented speech segment, the value of log $g(x|\lambda_{correct})$ will be low and the value of log $g_{Anti\text{-}Model}(x)$ will be high. This leads to low value of $G_{veri}(x)$. When $G_{veri}(x)$ is lower than a pre-defined threshold, speech segment x will be regarded as an accented speech segment.

B. AA-based accented speech segment verification

The second stage of accented speech verification is based on articulatory attributes (AAs). The basic assumption of the proposed articulatory attribute-based accented speech segment verification is that if a recognized phone has low confidence, there are two possibilities: 1) If the phone is normally pronounced but misrecognized, the articulatory attributes of this phone are similar to the normal pronunciation of the phone. 2) If the phone has accent and thus has low confidence, the articulatory attributes of this phone may be different from that of the normal pronunciation. Based on this assumption, if a specific accented speech has different AAs from its normal pronunciation, the candidate speech segment is regarded as accented. The articulatory attribute detector, trained using a speech database with articulatory attribute labels, estimates the attribute likelihood of a senone segment. All of these attribute likelihoods of a senone segment can be formed as an $M\times1$ AA vector, where M is the number of AAs. In this study, 14 AAs listed in Table 1 are used.

The values of the AA vector are the average likelihood of all the frames in a specific senone segment. Fig. 2 shows the AA-based accented speech verification process.

The attribute detector is constructed using an Artificial Neural Network (ANN). Equation (3) is used to estimate the likelihood of an articulatory attribute E for the speech frame corresponding to a senone of the phone HMM.

$$p(y = E \mid X_i^S) = T^{-1} \sum_{t=1}^{T} \left(\frac{\exp(w_E^T z(t))}{\sum_{j=1}^{M} \exp(w_j^T z(t))} \right) \quad (3)$$

$$z(t) = [1 \quad x(t) \quad u(t)]^T \quad (4)$$

$$u(t+1) = (1 + \exp(-v_E z(t)))^{-1} \quad (5)$$

where $z(t)$ is the input vector of a frame and composed of feature vector and the current state vector u(t) in the attribute detector. w_E and v_E are the weighting parameters for the output and the next state of the attribute detector, respectively. T is the total frame numbers in the candidate accented senone segment. An accented senone verifier characterized by a binary support vector machine (SVM) is proposed for accented speech verification. If a segment is classified as positive, the segment is regarded as non-accented under the assumption that a specific senone may have different AAs for the accented speech compared to the corresponding normal pronunciation. On the other hand, a segment is regarded as accented if it is classified as negative using the binary SVM.

TABLE I
14 ARTICULATORY ATTRIBUTES

Articulatory Attributes	
Anterior	*Nasal*
Back	*Round*
Consonantal	*Silence*
Continuant	*Strident*
Coronal	*Tense*
High	*Vocalic*
Low	*Voice*

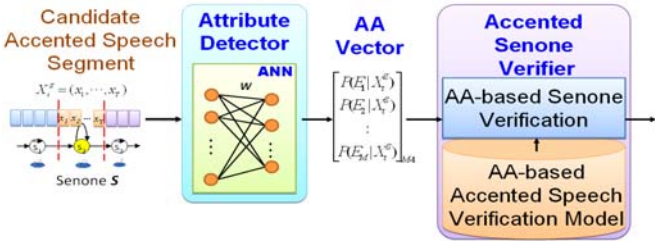

Fig. 2 AA-based accented speech verification process.

III. ACCENTED SPEECH MODEL GENERATION

After accented speech segment verification, accented speech data are used to build the acoustic models of the accented speech. The normal pronunciation and the corresponding accented speech of a specific senone are selected from the corpus and paired for transformation function construction. Nevertheless, accented acoustic model construction generally suffers from the problem of data sparseness. Thus, the accented speech segments with similar acoustic properties and articulatory attributes are clustered using a decision tree. The paired normal pronunciations and accented speech segments in the same leaf node of the decision tree are used to construct a linear transformation function to convert the normal models to an accented speech model.

A. Linear transformation function

Assume a specific accented speech and its normal pronunciation has a linear relationship. The linear transformation can be denoted as Y = AX + R, where X = (x_1, x_2, …, x_N), where N is the number of frames, X is a normal senone model, A is the transformation matrix, R is the transformation error, and Y is the accented senone with Y = (y_1, y_2, …, y_N). The coefficients A and R can be estimated by maximizing P(X,Y|λ), where λ includes HMM parameters and the values of A and R in the transformation function. P(X,Y|λ) is estimated as (6).

$$P(\mathbf{X},\mathbf{Y}|\lambda)=\sum_{\forall q}P(\mathbf{X},\mathbf{Y},q|\lambda)=\sum_{\forall q}\pi_{q_0}\prod_{n=1}^{N}a_{q_{n-1}q_n}b_{q_n}(\mathbf{x}_n,\mathbf{y}_n) \tag{6}$$

where

$$b_j(\mathbf{x}_n,\mathbf{y}_n)=b_j(\mathbf{y}_n|\mathbf{x}_n)b_j(\mathbf{x}_n) \tag{7}$$

$$b_j(\mathbf{y}_n|\mathbf{x}_n)=N(\mathbf{y}_n;\mathbf{A}_j\mathbf{x}_n+\mathbf{R}_j,\mathbf{\Sigma}_j^{\mathbf{y}}) \tag{8}$$

$$b_j(\mathbf{x}_n)=N(\mathbf{x}_n;\boldsymbol{\mu}_j^{x},\mathbf{\Sigma}_j^{x}) \tag{9}$$

Note that π is the initial probability, $\boldsymbol{a}$ is transition probability, $\boldsymbol{b}$ is output probability, $\boldsymbol{q}$ means the state, j denotes the state index and Σ is the variance matrix.. The EM algorithm and a Lagrange multiplier are applied to estimate the values of A and R. Finally, the values of A and R can be obtained from (10) and (11)

$$\mathbf{A}_j'=\left(\sum_{n=1}^{N}r_n(j)(\mathbf{y}_n-\mathbf{R}_j)\mathbf{x}_n^{T}\right)\left(\sum_{n=1}^{N}r_n(j)\mathbf{x}_n\mathbf{x}_n^{T}\right)^{-1} \tag{10}$$

$$\mathbf{R}_j'=\frac{\sum_{n=1}^{N}r_n(j)(\mathbf{y}_n-\mathbf{A}_j'\mathbf{x}_n)}{\sum_{n=1}^{N}r_n(j)} \tag{11}$$

where

$$\tilde{\pi}_i=\frac{\gamma_n(\boldsymbol{i})}{\sum_{i=1}^{\boldsymbol{M}}\gamma_n(\boldsymbol{i})}=\gamma_n(\boldsymbol{i}) \tag{12}$$

B. Transformation function selection using a decision tree

In this study, the paired normal pronunciation and accented speech segments along with their corresponding transformation functions are classified through a decision tree. The question set of the decision tree contains articulation-related questions, for instance, "the pronunciation of the model is liquid or not?" The paired speech data with similar articulatory attributes are expected to be classified into the same leaf node.

The node-splitting criterion is defined based on the generation error. If a split happens, the generation error from the models in two child nodes should be lower than the generation error from their parent node. The generation error in each leaf node can be defined as (13).

$$GenErr=\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\left\|y_m-(A_{m,t}x_m+R_{m,t})\right\|^2 \tag{13}$$

where y_m is the m-th target model (accented speech model) in a leaf node, x_m is the paired normal model, $A_{m,t}x_m+R_{m,t}$ (t=1,…T) is the transformation results generated from the t-th nearest linear transformation function. The T linear transformation functions with their corresponding normal models nearest to normal model x_m are selected to generate the transformation results, M is the total number of models in the node. The reason why we take T nearest linear transformation functions into consideration is that we expect the differences among transformation results generated by the T-best transformation functions should be smallest. Finally, normal models and the corresponding linear transformation functions are stored in the leaf nodes. Normal senones which have the same articulatory attributes and similar acoustic feature will be classified into the same leaf node. In order to achieve the best split, the generation error should be minimized. The generation error reduction (GE-Reduction) is computed as (14).

$$GE-Reduction=GenErr_p-\sum_i\frac{M_i}{M_p}GenErr_i \tag{14}$$

where $GenErr_p$ is the generation error of the parent node and $GenErr_i$ is the generation error of the i-th child node. M_i is the amount of data in the i-th child node and M_p is the amount of data in the parent node.

The linear transformation functions for each normal pronunciation and the corresponding accented pronunciation are stored in a leaf node. A normal pronunciation senone without a normal pronunciation and its corresponding accented pronunciation in the training corpus can select a specific transformation function based on the articulatory features and acoustic features. The input data of the decision tree is a Gaussian mixture of a normal senone model which contains 39 dimensional MFCC features. When the

transformation function is selected for a specific normal model, the input normal model starts from the root, then goes to the child node according to the answer of the articulation-related question in each node. Once reaching a leaf node, a pre-stored normal phone's transformation function can be selected according to the acoustic similarity. The selected transformation function is then used to transform the input normal model to generate a mixture of the corresponding accented speech model. A new accented speech model is obtained from the new generated mixtures.

IV. MODEL DISCRIMINATION VERIFICATION

Transformation functions and a decision tree are used to generate the potential accented speech models. But not all of them are robust for speech recognition. Models with low discriminability will degrade the recognition accuracy. Recognition complexity increases if too many models are used for recognition. This paper adopts a discrimination function to verify if a model is discriminative for recognition. Low discriminative models should be removed to avoid ambiguity.

The discrimination degree of a generated accented phone model is estimated by (15)

$$d_i^{accent\_N} = -g(Y_i \mid \lambda_{Y_i}) + \max_{m \in X} \{g(Y_i \mid \lambda_m)\} \tag{15}$$

where $d_i^{accent\_N}$ is the discrimination degree of the *i*-th generated accented phone model and $g(\ )$ is a similarity measure function. λ_{Y_i} is the *i*-th generated accented phone model. Y_i denotes the mean and variance of the *i*-th generated accented phone model. λ_m is the *m*-th normal model. The lower the value of $d_i^{accent\_N}$, the higher the discriminability of the generated accented phone model. On the other hand, a dynamic discrimination threshold of the *i*-th generated accented phone model to its corresponding normal phone model is defined in (16).

$$d_i^{normal\_N} = -g(Y_i \mid \lambda_{X_i}) + \max_{m \in X, m \neq X_i} \{g(Y_i \mid \lambda_m)\} \tag{16}$$

where λ_{X_i} is the original normal phone model with respect to the accented data Y_i. λ_m represents all the normal phone models except λ_{X_i}. If the value of $d_i^{accent\_N}$ is lower than that of $d_i^{normal\_N}$, the generated accented phone model will be more discriminative than the original normal phone model. Hence, this model is regarded as an accented phone model.

V. EXPERIMENTAL RESULTS

In this study, TIMIT database was used to train the English acoustic models. The numbers of sentences for males and females are 3,020 and 1,280, respectively. The goal of this study is to design a robust ASR system which can eliminate the recognition accuracy degradation problem due to accents produced by non-native English speakers. Hence, the EAT (English Across Taiwan) corpus was applied for evaluation. EAT contains English sentences spoken by Taiwanese people. The training sentences of EAT corpus are divided into two parts. One is composed of English sentences spoken by the students majoring in English. These sentences are manually selected and used as normally pronounced sentences. This part contains 194 and 256 sentences for males and females, respectively. In order to reduce the mismatch between TIMIT and EAT, these normally pronounced sentences were used to adapt the TIMIT-trained models based on MLLR. In the following experiments, the adapted models are named as normal phone models. The other part of EAT contains English sentences which were spoken by the students not majoring in English. These sentences are manually selected and regarded as accented speech. This part contains 241 and 209 speech utterances from males and females, respectively. 10-fold cross-validation experiments were conducted on English speech recognition tests. In each validation, 45 sentences were selected as the test data and the other 405 sentences were used as the training data. 39-dimensional MFCCs were used as the speech features. For constructing an ASR system, HTK (Hidden Markov Model) toolkit was applied to train the acoustic models for speech recognition. The acoustic models are tri-phone models. Each phone model contains three states with 16 Gaussian mixtures. The total number of defined English phones is 40. In the ASR system, Mandarin phone models and English phone models are constructed to form a bilingual model set. However, in order to focus on English accented speech recognition, the Mandarin recognition results are not reported in this paper. The following English speech recognition results come from this bilingual ASR system, which combines English and Mandarin phone models together. The language model was built by using the sentences of TCC300, TIMIT, EAT and Chinese Gigaword which contains 1.1 billion Chinese characters.

In this study, the final acoustic phone models including normal phone models and the generated accented phone models are called expanded models. The experiment compared the recognition accuracy of using the normal models and the expanded models. Fig. 3 shows the evaluation results of using normal phone models and expanded phone models in recognizing the EAT corpus. In recognizing accented speech, the expanded models outperform the normal models. Using expanded models achieved 71.7% word accuracy higher than 68.6% achieved by using normal models in recognizing the accented speech data. In recognizing normal speech, the expanded phone models also outperform traditional normal phone models. The results reveal that the proposed expanded phone models not only can improve the accuracy in recognizing accented English speech but also the robustness in recognizing normal English speech. Furthermore, Fig. 4 shows the accuracy in recognizing normal and accented phones in EAT by using the expanded models and normal models. In this experiment, the phone accuracies

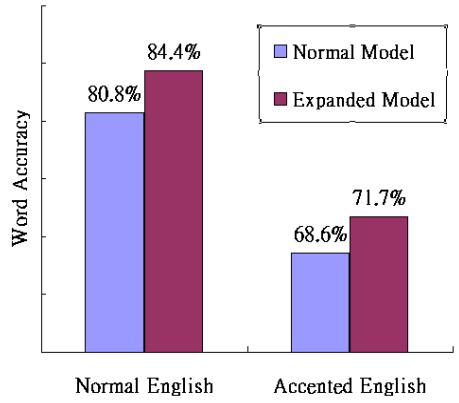

Fig. 3 Evaluation results of normal phone models and expanded phone models

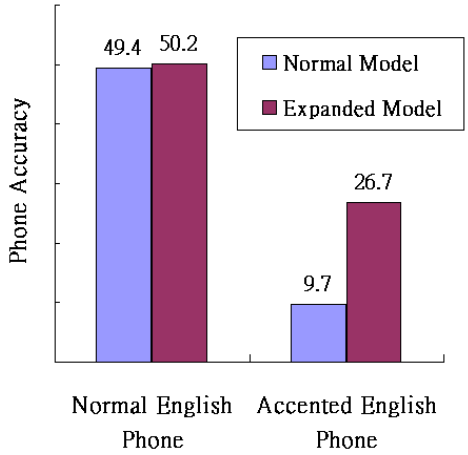

Fig. 4 Average cross validation results of the recognition performance for normal phone models and expanded phone models in recognizing normal and accented phones

of using normal phone models and expanded phone models in recognizing normal and accented speech phones are presented. The experiment is also conducted through 10-fold cross validation. Before the experiment, we have annotated whether a specific phone segment is an accented segment or not manually. After the experiment, the recognition accuracies of the accented speech data and normal speech data were calculated, respectively. Form the result, the proposed expanded phone models can outperform the traditional normal phone models in recognizing both accented and normal speech. The accuracy of using expanded phone models achieved 26.7% phone recognition rate in accented phone recognition, which is 17.0% higher than that using normal phone models. From the results, we can conclude that the proposed method not only obtains a satisfactory performance in recognizing normal phones, but also significantly improves the accuracy in recognizing accented speech. In addition, we also conducted the experiment on the comparison of the proposed transformation-based method and MLLR-based method which uses 8 regression classes for accented phone model construction. The word accuracy from the models adapted by the sentences pronounced by the non-major English students achieved 69.5% lower than 71.7% for transformation-based method. This indicates that the proposed transformation-based method outperforms MLLR-based method on accented model construction for non-native speech recognition.

VI. Conclusions

This study focuses on generating accented acoustic models for recognition of non-native speech with accents. In accented speech segment extraction, acoustic features and articulatory attributes are considered. For the extracted accented speech segment, the linear transformation functions between normal phones and accented phones are constructed. A decision tree is constructed based on the articulatory attributes and acoustic features and used to select a suitable transformation function to generate accented models with unseen data. Finally, discrimination verification is performed to remove the accented models with low discriminability.

The experimental results reveal that the proposed ASR system using expanded models can improve the ability of recognizing accented speech while retain good accuracy in recognizing normal speech. The recognition rates of the proposed system achieved improvements of 3.1% compared to those using traditional ASR in recognizing English speech from non-native speakers. On the other hand, the accuracy of recognizing accented phones is 17.0% higher than that using normal phone models. Finally, the proposed method also outperforms MLLR-based method by 2.2% in recognizing accented speech.

References

[1] L.M. Arslan and J. H. L. Hansen, "Language accent classification in American English," *IEEE Trans. Speech Communication*, Vol. 18, No. 4, pp. 353-367, 1996.

[2] K. Kulkarni, S. Sengupta, V. Ramasubramanian, J. G. Bauer and G. Stemmer, "Accented Indian English ASR: Some Early Results," *IEEE Workshop on Spoken Language Technology*, India, 2008.

[3] Y.-Y. Pu, J. Yang, H. Wei and D. Xu, "A Study on Yunnan Dialectal Chinese Speech Recognition," in *Proc. of Seventh International Conference on Machine Learning and Cybernetics*, Kunming, 2008.

[4] Z. Wang, T. Schultz and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. of ICASSP*, pp. 540-543 ,2003..

[5] P. Fung and Y. Liu, "Effects and Modeling of Phonetic and Acoustic Confusions in Accented Speech Recognition," *Journal of the Acoustical Society of America*, Vol.118, Issue 5, pp.3279-3293, 2005.

[6] S.-M. Siniscalchi, T. Svendsen and C.-H. Lee, "Toward A Detector-Based Universal Phone Recognizer," In *Proc. of ICASSP*, 4261-4264, 2008.