Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 54 (2012) 40–54

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic analysis of Mandarin accented English using phonological features

Abhijeet Sangwan, John H.L. Hansen *

Center for Robust Speech Systems (CRSS), University of Texas at Dallas (UTD), Richardson, TX, USA

Received 14 August 2009; received in revised form 26 April 2011; accepted 16 June 2011
Available online 30 June 2011

Abstract

The problem of accent analysis and modeling has been considered from a variety of domains, including linguistic structure, statistical analysis of speech production features, and HMM/GMM (hidden Markov model/Gaussian mixture model) model classification. These studies however fail to connect speech production from a temporal perspective through a final classification strategy. Here, a novel accent analysis system and methodology which exploits the power of phonological features (PFs) is presented. The proposed system exploits the knowledge of articulation embedded in phonology by building Markov models (MMs) of PFs extracted from accented speech. The Markov models capture information in the PF space along two dimensions of articulation: PF state-transitions and state-durations. Furthermore, by utilizing MMs of native and non-native accents, a new statistical measure of “accentedness” is developed which rates the articulation of a word by a speaker on a scale of native-like (+1) to non-native like (−1). The proposed methodology is then used to perform an automatic cross-sectional study of accented English spoken by native speakers of Mandarin Chinese (N-MC). The experimental results demonstrate the capability of the proposed system to perform quantitative as well as qualitative analysis of foreign accents. The work developed in this study can be easily expanded into language learning systems, and has potential impact in the areas of speaker recognition and ASR (automatic speech recognition).

Keywords: Phonological features; Accent analysis; Non-native speaker traits

1. Introduction

Automatic accent analysis and classification is useful in speech science, with impact in many areas of speech technology such as automatic speech recognition (ASR) (Salvi, 2003; Zheng et al., 2005), speaker recognition (Mangayyagar i et al., 2008), pronunciation modeling, pronunciation scoring, and language learning (Mak et al., 2003; Neri et al., 2006; Wei et al., 2006). Accent analysis is the process of identifying speech characteristics that contribute to a speaker’s accent. Accent structure can be based on one of three perspectives: (i) physical speech production analysis including phonemic, prosody, and linguistic structure, (ii) acoustic waveform analysis based on signal processing feature extraction, and (iii) human perception, which is based on the salient traits extracted by the listener which characterize an accent. These represent the science and technology domains for accent research. Alternatively, accent classification identifies a speaker’s accent based on the most discriminating speech characteristics. Here, cepstrum based features are most widely used for accent classification (Angkititrakul and Hansen, 2006; Choueiter et al., 2008). Additionally, low level speech features such as VOT (voice-onset time), word/phone durations, intonation patterns, formant-behavior, *etc.* have also been used for modeling and classifying accents (Arslan and Hansen, 1996a,b; Das and Hansen, 2004; Hansen et al., 2010). Finally, a number of modeling techniques including GMMs (Gaussian mixture models), HMMs (hidden Markov models), and SVMs (support vector machines) have been employed to learn accent characteristics and have

* Corresponding author.
E-mail address: john.hansen@utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.06.003

shown good performance in classification (Arslan and Hansen, 1996b; Pedersen and Diederich, 2007). While the above-mentioned features and modeling techniques provide good classification accuracy, they do not offer a comprehensive insight into the major differences among the accents under consideration. Since the origins of accent are embedded in production differences, it would be beneficial to automatically capture, compare and contrast the major articulatory characteristics of accents. Therefore, it is asserted that a phonological features (PFs) based framework would offer the necessary breadth and depth to comprehensively analyze distinct accents. Herein, the ability of PFs to capture fine articulatory variations in speech has been demonstrated in the research literature (Scharenborg et al., 2007; Sangwan and Hansen, 2008; King et al., 2007). This motivates the design and development of a PF-based accent analysis system.

In this study, the objective is to develop an accent analysis system that automatically models the major differences in articulation characteristics of two accent groups (native and non-native speakers). Using these accent models, the system is able to identify the speech characteristics of an individual speaker as native or non-native. In fact, the models are not only used to identify speech characteristics of different accents but to score them as well. These new scores represent the measure of "accentedness". As shown in Fig. 1, it is proposed to form a continuation of "accentedness", which is bounded over $[-1, +1]$ where a value of -1 and $+1$ imply extremely non-native-like and native-like accent characteristics. The bounds identify the extremes in proficiency, where an individual speakers proficiency can be rated on this continuum. It is expected that the distribution of accentedness scores of non-native and native speakers would be similar to that shown in Fig. 1(a).

Our approach towards building the automatic accent analysis system relies on the use of phonological features (PFs). The use of PFs is especially beneficial, owing to the close relationship between PFs and articulatory/acoustic phonetics. The various PF dimensions enable a comprehensive accent analysis in the space of speech articulators. In this manner, the proposed accent analysis system is able to assign accentedness scores to various articulatory/acoustic traits of a speaker (*e.g.*, aspiration, nasalization, rounding *etc.*). Therefore, PFs allow the proposed system to independently look at fine articulatory/acoustic details resulting in a more refined assessment of accent characteristics. Alternatively, PFs also enable differential diagnosis of a speaker's accent as shown in Fig. 1(c). As seen in the figure, the different PF dimensions can be independently assessed for accentedness per speaker resulting in unique accent profiles. These accent profiles can provide very useful information for language learners as they clearly identify the students strengths and weaknesses. Furthermore, as shown in Fig. 1(b), a longitudinal study of the accent profile would identify areas of speech production improvement and stagnation as accent relaying properties. Finally, the information provided by accent profiles can also be potentially used as (i) input features for dialect, accent, speaker or language identification systems, and (ii) conditioning knowledge for ASR systems.

The proposed analysis system models accent by exploiting two important properties of articulation embedded within PF sequences, namely, state-occupancy and state-transitions. State-occupancy captures the durational aspect of articulation. State-transitions capture the characteristics of articulatory motion. For example, the tongue movement from a velar to dental place-of-articulation is a state-transition. Alternatively, the duration spent in the dental place-of-articulation is state-duration. Here, it is hypothesized that given the same articulation task (*e.g.*, pronouncing a word) the statistical nature of state-transitions and state-durations of native and non-native speakers would be dramatically different. Hence, we propose to learn the statistical nature of state-transitions and state-durations for native and non-native articulation using Markov models (MMs). Subsequently, the MMs are used to compute the likelihood that an utterance was articulated by a native or non-native speaker. In this manner, the likelihoods can then be used to generate accentedness scores.

The proposed accent analysis system is evaluated on the CU-Accent corpus (Angkititrakul and Hansen, 2006). In our experiments, native speakers of American English (N-AE) and Mandarin Chinese (N-MC) are drawn from CU-Accent as the native and non-native speaker groups. The N-MC speakers are further divided into two groups based on their AE-exposure (which is assumed to be equal to their stay in U.S.A). The N-MC 1 and N-MC 2 groups correspond to the low and high exposure N-MC speakers, respectively. Using the target speakers in the above-mentioned speaker-groups, a number of experiments are conducted to demonstrate the accuracy and utility of the proposed system. In the first experiment, human-assigned accentedness scores are collected for the N-AE and N-MC speakers in a listener evaluation study. Subsequently, as shown in Fig. 1(d) the correlation between human-assigned scores and automatic scores (generated by the proposed system) is studied. Our results show a good correlation (0.8, $p < 0.0001$) between the human-assigned and machine-assigned accentedness scores. Additionally, the correlation between the machine-assigned accentedness scores and L2-exposure of N-MC speakers is also reported. The results reported in this study corroborate previous findings where non-native proficiency is seen to increase with L2-exposure (Flege et al., 1997; Flege, 1988; Jia et al., 2006). Encouraged by these findings, an in-depth differential analysis which compares and contrasts the articulatory dissimilarities of the native and non-native speaker groups is performed. As shown in Fig. 1(c), the differential analysis assigns accentedness scores to every PF-dimension. Hence, the proficiency of speakers is easily compared along individual articulators. The differential analysis performed in this study suggests an imbalanced increase in proficiency among N-MC speakers with increased L2-exposure. Particularly, it is observed that N-MC speakers

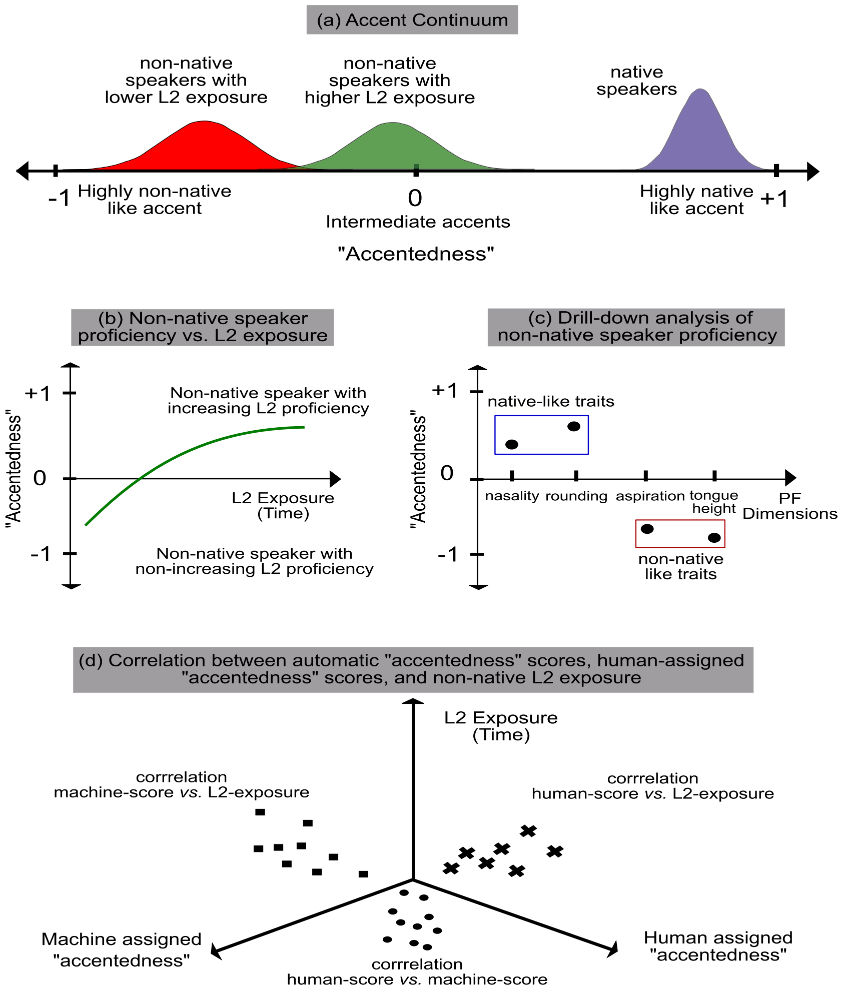
(a) Accent Continuum
non-native speakers with lower L2 exposure
non-native speakers with higher L2 exposure
native speakers
-1 Highly non-native like accent
0 Intermediate accents
Highly native like accent +1
"Accentedness"
(b) Non-native speaker proficiency vs. L2 exposure
+1
0
-1
"Accentedness"
Non-native speaker with increasing L2 proficiency
L2 Exposure (Time)
Non-native speaker with non-increasing L2 proficiency
(c) Drill-down analysis of non-native speaker proficiency
native-like traits
nasality rounding aspiration tongue height
PF Dimensions
non-native like traits
(d) Correlation between automatic "accentedness" scores, human-assigned "accentedness" scores, and non-native L2 exposure
L2 Exposure (Time)
corrrelation machine-score vs. L2-exposure
corrrelation human-score vs. L2-exposure
Machine assigned "accentedness"
corrrelation human-score vs. machine-score
Human assigned "accentedness"

Fig. 1. (a) The proposed measure of accent proficiency ("accentedness") on a scale of −1-to-+1 where −1 and 1 correspond to extremely non-native and native-like accent characteristics. (b) The "accentedness" measure can be used to track the progress of a student's proficiency over time (where a non-native speaker may "level-off" at some peak in "nativeness"), and (c) perform an in-depth differential analysis across the articulator space. (d) In the proposed study, the efficacy of the machine-assigned "accentedness" is determined by a comparison to the human-assigned accent scores as well as the L2-exposure of non-native speakers.

gain greater proficiency in (i) vowel articulation as opposed to consonant articulation, and (ii) duration aspects of articulation as opposed to transitional aspects. In this manner, the proposed accent analysis system is able to offer a comprehensive comparative analysis of non-native and native articulation. Additionally, the proposed accent-analysis algorithm is easy to implement, and versatile in its usage. Therefore, the proposed system is beneficial to language learners, and speech scientists as an assistive analysis tool. Finally, the proposed scheme can be further developed to integrate into speech technology such as ASR, speaker recognition, and accent/dialect/language identification.

The remainder of this paper is organized as follows: in Section 2, the CU-Accent speech corpus is described. In this study, data from CU-Accent corpus has been used for development and analysis. In Section 3, a brief review of PFs with respect to speech technology is presented. The hybrid features (HFs) system is also introduced, and our HMM-based PF extraction system is described. In Section 4, the proposed accent analysis model based on Markov models is developed. Finally, in Section 5, the in-depth comparative analysis of native *vs.* non-native accent is presented and discussed.

2. CU-accent speech corpus

The CU-Accent corpus consists of speech utterances spoken by native speakers of American English (AE), Mandarin Chinese (MC), Turkish, Thai, Spanish, German, Japanese, Hindi, and French speakers (Angkititrakul and Hansen, 2006). The corpus consists of several male as well as female speakers per native-language (as mentioned above) where the utterances for each speaker were recorded over multiple sessions. In each data-collection session, the speakers were required to speak 5 tokens of 23 isolated words, and 4 phrases in English as well as their native-tongue; along with 1 min of spontaneous speech on any topic of their choice. The lists of prompted phrases and isolated words are shown in Table 3. The choice of words and

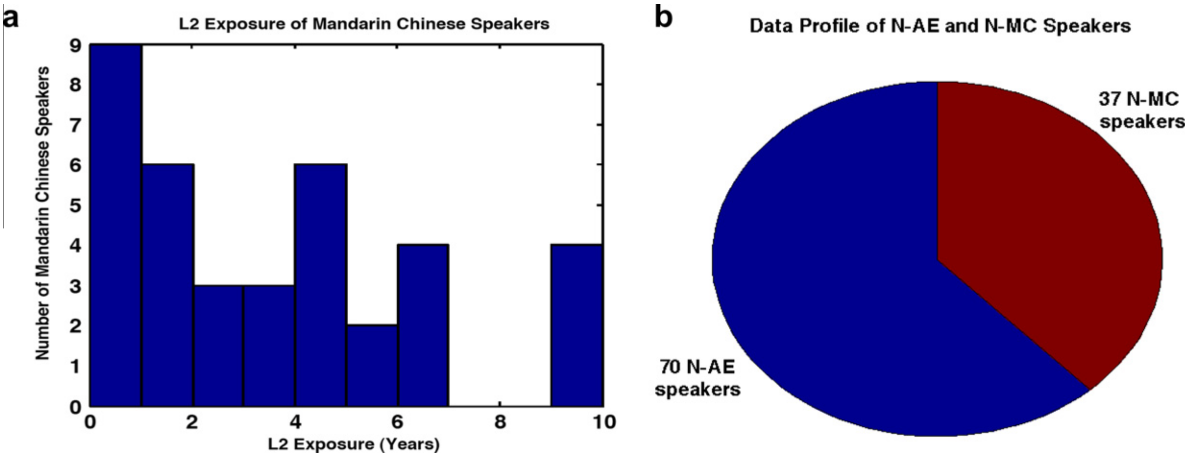

Fig. 2. Profile of speakers chosen for the accent study: (a) Distribution of N-MC speakers with respect to their L2 exposure, and (b) CU-Accent data profile used for this study.

phrases ensured that phonetic-composition and phonetic-transitions which are known to be problematic for non-native speakers of AE existed in the corpus (Chreist, 1964). The speech was collected over a telephone channel, digitized at 8kHz, and stored in 16-bit linear PCM (pulse coded modulation) format.

In our study, we have used all native speakers of American English (AE) and Mandarin Chinese (MC). The AE and MC speakers are the two largest linguistic groups in the corpus which provides sufficient diversity for robust modeling and analysis. As shown in Fig. 2(b), a total of 37 N-MC and 60 N-AE speakers are available in the CU-Accent corpus and have been used in this study. Since each speaker participated in 5 sessions of data collection, a total of 8050 N-AE (60speakers × 23words × 5sessions) and 4255 N-MC (37speakers × 23words × 5sessions) data samples of isolated word utterances are available for modeling and analysis. Additionally, Fig. 2(a) also shows the L2 exposure (American English) of MC speakers. It is seen that the corpus consists of MC speakers with 0–10 yr of exposure to AE.

3. Phonological features (PFs)

Phonological features (PFs) are a generic concept in linguistics with several manifestations such as the binary features in Sound-Patterns in English (SPE), government phonology (GP), multi-valued (MV) features, and hybrid features (HFs) (Scharenborg et al., 2007; Frankel et al., 2007a). The different PFs definitions are inspired by articulatory, acoustical, phonological, or a combination of different aspects of speech. In this paper, we use the HFs definition owing to their close relationship to articulatory phonetics.

3.1. Phonological features in speech technology

Within speech technology, phonological features (PFs) have been generally used for speech recognition. The use of PFs in ASR (automatic speech recognition) is motivated by the ability of PFs to address production-related issues. PF research for speech recognition has focussed on two aspects: (i) PF extraction which estimates the underlying phonology in speech signals, and (ii) PF integration which leverages the new phonological knowledge to improve speech system performance, especially speech recognition. Speech recognition tasks that incorporate PFs have been largely restricted to phone recognition. However, good improvements in WERs (word error rates) have been observed in systems where PF knowledge has been integrated with standard ASR to form hybrid systems. Herein, the fusion of a traditional ASR stream with a PF knowledge stream has been performed at the acoustic, decoding, and/or lattice levels (Metze and Waibel, 2002).

A wide range of modeling techniques such as ANNs (artificial neural networks), HMMs (hidden Markov models), GMMs (Gaussian mixture models), SVMs (support vector machines) have been successfully employed for extracting PFs from speech (King and Taylor, 2000; Scharenborg et al., 2007). A comparative study of the extraction strategies performed by Markov et al. Markov et al. (2006), indicates that the performance of HMM is comparable to other machine learning techniques, whereas a hybrid extraction strategy combining the best systems might yield higher performance. Since individual PF streams for a speech signal are correlated, several schemes such as DBNs (Dynamic Bayesian Networks), and CRFs (Conditional Random Fields) have been proposed as a means of exploiting this redundancy to improve PF extraction quality, and consequently ASR performance (Frankel et al., 2007b; Morris and Fosler-Lussier, 2008).

Recently, PFs have been successfully employed in several speech applications such as whispered speech recognition, speaker verification, pronunciation modeling, speaker adaptation, and variable frame rate (VFR) based speech recognition front-end (Metze, 2007; Leung et al., 2006; Jou et al., 2005; Tepperman and Narayanan, 2008; Sangwan and Hansen, 2007). The promising results obtained by using PFs in these new applications indicate the broader impact that PFs can have in speech technology. Among the above mentioned topics in speech processing, the use of PFs in accent/dialect classification or analysis systems has received only limited attention. Given findings that indicate PFs ability to capture fine articulatory production variations, accent analysis and classification based on a PF framework is highly intuitive.

3.2. Hybrid features (HFs)

As shown in Table 1, HFs (hybrid features) offer a comprehensive set of PF dimensions that are largely inspired by articulatory phonetics. The PF dimensions within the HF system cover sufficient breadth and depth in phonology to simultaneously allow meaningful detection accuracy as well as improved analysis capability. HFs were predominantly designed to assist in engineering applications such as ASR which explains some redundancy in the design. The HF features were first introduced in (Frankel et al., 2007a), and the different PF dimensions along with their closed value sets are shown in Table 1. While place, degree, nasality, rounding, and glottal serve mostly towards describing consonants characteristics; vowel, height, and frontness deal exclusively with vowels. The entire HF set can be found in (Frankel et al., 2007a).

In what follows, the HF based PF detection system used in our analysis is described. Our HF extraction scheme is based on an HMM framework. In our extraction system, all speech utterances are pre-emphasized with a factor of 0.97, and subsequent frame analysis using 25ms windows with a 15ms overlap. Thereafter, 13 dimensional MFCC (Mel frequency cepstral coefficient) vectors are extracted using a set of 40 triangular filters to simulate the Mel-scale, along with their delta and delta-delta MFCCs concatenated to the static vector to form a 39-dimensional feature vector. Cepstral mean subtraction (CMS) and automatic gain control (AGC) are also employed as part of the overall system. The features are used to train an HMM (hidden Markov model) based classification system. Context independent modeling with diagonal covariance matrices are used to model the evolution of the PFs in the signal. Furthermore, the HMM topology used for modeling is a 3-state left-to-right model with no state skipping. The training transcriptions for each HF type is directly obtained by mapping the phone sequence for each sentence to its equivalent HF sequence using a pre-defined map (King and Taylor, 2000). In our experiments, a simple bigram phonotactic language model is trained for each HF type and used in the decoding of the HF sequence for each test utterance. All detection experiments are performed using the SPHINX recognition system. The performance of the HF detection system on the TIMIT corpus (test only) in terms of frame level accuracy is summarized in Table 2. It is noted that phone to PFs mapping was employed to establish the frame level ground truth, since actual PF values are not available for the TIMIT corpus. The above HF detection system was trained on speech data of N-AE speakers obtained from the CU-Accent and TIMIT corpora (Garofolo et al., 1993).

Table 2
Frame-level accuracy of hybrid features (HFs) detection.

Hybrid feature	Frame level accuracy (%)
Place	80.26
Degree	79.24
Nasality	93.28
Rounding	81.55
Glottal	89.89
Vowel	71.01
Height	83.79
Frontness	83.92

4. Proposed accent model

In this section, we develop the proposed accent model. The HF system described in Section 3.2 captures two important aspects of articulation, namely, the HF state-transitions as well as HF state-occupancy. For example, in the articulation of the diphthong /aw/, the tongue shifts from a low-to-high position while occupying each state for a finite duration of time. As shown in Fig. 3, the vertical and horizontal movements of the tongue from a low to high position, and mid-front to mid-back position would be adequately captured by the HF-types: height and frontness, respectively. This change in the value of HF-type is

Table 1
Hybrid features (HFs): dimensions and the corresponding value-sets.

PF dimension	Cardinality	Values
Place	10	Silence, none, labial, alveolar, post-alveolar, dental,labio-dental, velar, lateral, rhotic
Degree	6	Silence, vowel, closure, fricative, flap, approximant
Nasality	3	Silence, +, −
Rounding	3	Silence, +, −
Glottal	4	Silence, voiced, voiceless, aspirated
Vowel	24	Silence, aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, nil, eh, er,ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw
Height	8	Silence, low, mid, mid-low, high, nil, mid-high, very-high
Frontness	7	Silence, back, mid, mid-front, mid-back, front, nil

Table 3
Composition of the CU-Accent corpus.

Nature of speech data	Total unique tokens per speaker	Repetitions per token per speaker	Transcription
Isolated words	23	5	Aluminum, bird, boy, bringing, Catch, change, communication, feet, Hear, line, look, pump, root, south, Student, target, teeth, there, thirty, Three, voice, white, would
Phrases	8	4	This is my mother. He took my book. How old are you ? Where are you going ?
Spontaneous	1	1	1 min long monologue which varied from speaker to speaker

referred to as state-transition (*e.g.*, the HF-type height moving from state low to high). It is easy to see that state-transitions capture the corresponding articulatory evolution of an uttered word (or sentence). Furthermore, as the HF-types move from one state to another they also persist in each state for a finite duration of time. This time duration spent in a state is referred to as state-occupancy and is measured in the number of frames.

We believe that state-transition and occupancy capture articulation information that must shed light on critical aspects of accented articulation. We propose the use of a Markov process to model state-transition and occupancy as described above. The formulation of the Markov process is evident from the illustration of the state-machines in Fig. 3, and is fully developed in Section 4.1. Since this study focusses on aspects of Mandarin accented English, we estimate the parameters of the Markov process separately for L1 (*i.e.*, English spoken by AE speakers) and L2 (*i.e.*, English spoken by MC speakers) groups. Herein, the L1 and L2 articulation models establish extrema within the phonological (or articulation) space where individual articulation strategies can be gauged for their "accentedness". Furthermore, an average comparison between the L1 and L2 models also reveals an overall picture of the major articulatory differences when the L1 and L2 groups are compared in general. We believe that some of these production differences must be perceivable and therefore significant indicators of accent.

4.1. Markov process

Articulation of a given word **W** involves production of a sequence of phones. In terms of HFs, articulation involves transitioning through a series of HF states. Let the *j*th state of the *q*th HF-type be denoted by S_{qj}. Furthermore, the speaker also occupies O_{qj} frames of speech in the HF state S_{qj}. Therefore, the articulation process can be conveniently captured by the ordered sequence of pairs,

$$\{(S_{q1}, O_{q1}), (S_{q2}, O_{q2}), \ldots, (S_{qN}, O_{qN})\}, \tag{1}$$

where the articulation is assumed to prolong over *N* states. The sequence of state-occupancy pairs forms a Markov

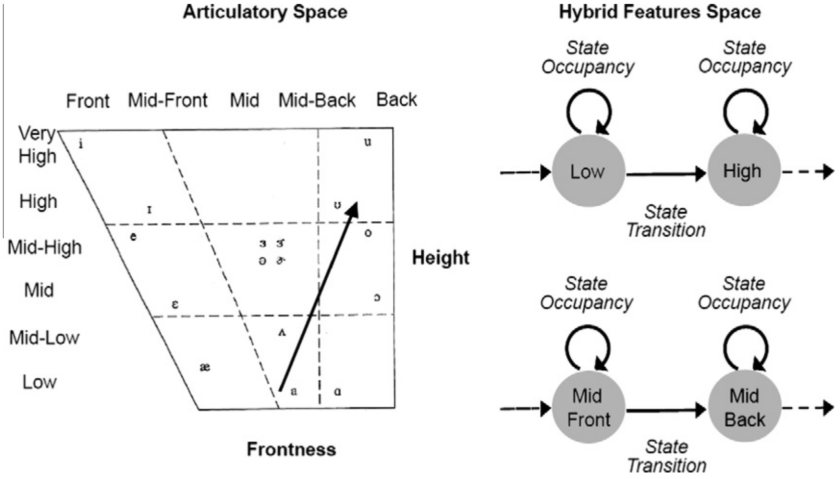

Fig. 3. The arrow indicates the direction of tongue movement in production of the diphthong /aw/. As the tongue moves from low-mid position to high-back position in the articulator space, the HFs 'height' and 'frontness' capture the motion in the corresponding PF space. As shown, the HFs capture two aspects of the movement: state-transitions and state-occupancies.

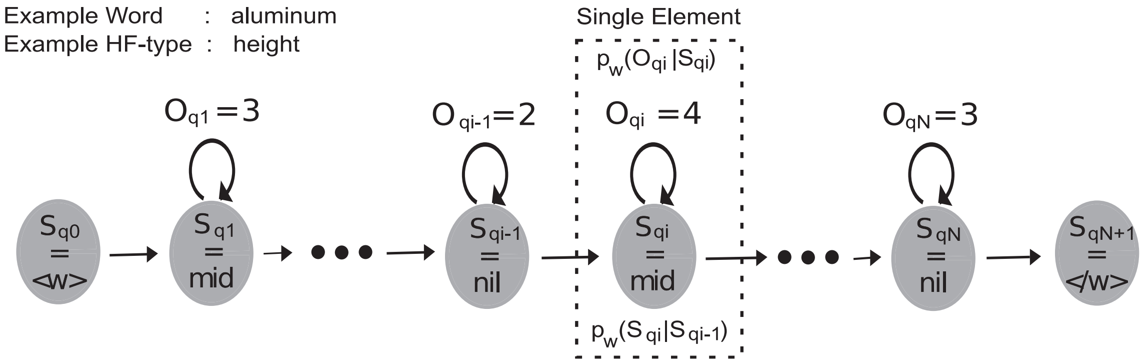

Fig. 4. The proposed accent model learns the articulatory evolution of a word (*e.g.* aluminum) along each HF-type (*e.g.* height) in terms of state-transitions and state-occupancies. The state-occupancies $p_{\mathbf{w}}(O_{qi}|S_{qi})$ are modeled by Gamma-distribution where the distribution parameters are estimated from data. The state-transitions probabilities $p_{\mathbf{w}}(S_{qi}|S_{qi-1})$ are also estimated from data.

process as shown in Fig. 4. Here, the relation between the production space and the proposed model space becomes clearer by comparing Figs. 3 and 4. The duration and nature of articulator movements is captured by the model states and state-transitions. Next, in order to model the entry and exit into the Markov process, we define ordered pairs $(S_{q0}, 0)$ and $(S_{qN+1}, 0)$ as the entry and exit states. As shown in Fig. 4, the entry and exit states always carry the values $\langle w\rangle$ and $\langle/w\rangle$. Furthermore, the occupancies of both entry and exit states are always 0. By including the entry and exit states, the expanded state-occupancy sequence is given by:

$$\{(S_{q0},0),(S_{q1},O_{q1}),(S_{q2},O_{q2}),\ldots,(S_{qN},O_{qN}),(S_{qN+1},0)\}. \tag{2}$$

Let the joint likelihood of observing the qth HF states and occupancies be given by:

$$\begin{aligned}\Lambda(\mathbf{W}_q) &= p_{\mathbf{w}}(O_{q0},\ldots,O_{qN},O_{qN+1},S_{q0},\ldots,S_{qN},S_{qN+1}),\\ &= p_{\mathbf{w}}(\mathbf{O}_q,\mathbf{S}_q),\end{aligned} \tag{3}$$

where $\mathbf{W}_q$ is the qth HF-sequence for the word $\mathbf{W}$, and

$$\mathbf{O}_q \equiv \{O_{q0},O_{q1},\ldots,O_{qN},O_{qN+1}\} \tag{4}$$

and

$$\mathbf{S}_q \equiv \{S_{q0},S_{q1},\ldots,S_{qN},S_{qN+1}\} \tag{5}$$

represent the series of occupancy and state observations respectively. Using conditional probability, Eq. (3) can be written as,

$$\Lambda(\mathbf{W}_q) = p_{\mathbf{w}}(\mathbf{O}_q|\mathbf{S}_q)p_{\mathbf{w}}(\mathbf{S}_q). \tag{6}$$

By assuming that the occupancy observations are mutually independent and depend only upon their respective states alone, the first part of Eq. (6) is given by:

$$p_{\mathbf{w}}(\mathbf{O}_q|\mathbf{S}_q) = \prod_{i=0}^{N+1} p_{\mathbf{w}}(O_{qi}|S_{qi}). \tag{7}$$

Here $p_{\mathbf{w}}(O_{qi}|S_{qi})$ is the probability of observing an occupancy of O_i frames given that the state occupied is S_i, and the word in question $\mathbf{W}$. It may be noted that here, we have $p_{\mathbf{w}}(O_{q0}|S_{q0}) = p_{\mathbf{w}}(O_{qN+1}|S_{qN+1}) = 1$. Furthermore, if the state transitions are assumed to be a Markov process (*i.e.*, the state transition into S_{qi} depends upon previous state S_{qi-1} alone), then the second part of Eq. (6) is given by:

$$p_{\mathbf{w}}(\mathbf{S}_q) = \prod_{i=1}^{N+1} p_{\mathbf{w}}(S_{qi}|S_{qi-1}), \tag{8}$$

where $p_{\mathbf{w}}(S_{qi}|S_{qi-1})$ is the transition probability due to the Markov process assumption. By combining Eqs. (7) and (8), we obtain the following expression for the joint likelihood in Eq. (3):

$$\Lambda(\mathbf{W}_q) = \prod_{i=1}^{N+1} p_{\mathbf{w}}(O_{qi}|S_{qi})p_{\mathbf{w}}(S_{qi}|S_{qi-1}). \tag{9}$$

Alternatively, the log-likelihood of jointly observing the qth HF states and occupancies are given by:

$$\log(\Lambda(\mathbf{W}_q)) = \sum_{i=1}^{N+1} \log(p_{\mathbf{w}}(O_{qi}|S_{qi})p_{\mathbf{w}}(S_{qi}|S_{qi-1})). \tag{10}$$

In our study, the distribution $p_{\mathbf{w}}(O_{qi}|S_{qi})$ is assumed to be Gamma distributed as *i.e.*,

$$p_{\mathbf{w}}(O_{qi}|S_{qi}) = (O_{qi})^{(\kappa_{\mathbf{w}S_{qi}}-1)}\frac{\exp\left(-\frac{O_{qi}}{\theta_{\mathbf{w}S_{qi}}}\right)}{(\theta_{\mathbf{w}S_{qi}})^{\kappa_{\mathbf{w}S_{qi}}}\Gamma(\kappa_{\mathbf{w}S_{qi}})}, \tag{11}$$

where $\kappa_{\mathbf{w}S_{qi}}$ and $\theta_{\mathbf{w}S_{qi}}$ are the shape and scale parameter of S_{qi} reflecting the HF-state and word $\mathbf{W}$, respectively. In our accent model, we use the ML (maximum likelihood) estimates of the distribution parameters that are directly computed from the data. Similarly, the ML estimates of transition probabilities $p_{\mathbf{w}}(S_{qi}|S_{qi-1})$ are determined directly from the data as well. As a result, for each word $\mathbf{W}$ and HF-state the number of state-occupancy densities $p_{\mathbf{w}}(O_{qi}|S_{qi})$ and state-transition densities $p_{\mathbf{w}}(S_{qi}|S_{qi-1})$ are V and V^2, where V is the number of values the qth HF-state can take. The set of values that HF-types can take are shown in Table 1. The above-mentioned development formalizes the proposed accent model. However, in order to directly compare and contrast the articulation characteristics of two accents, we proceed towards developing a differential model that simultaneously makes use of the $L1$ and $L2$ Markov process based accent models.

Let the Markov process based accent models for $L1$ and $L2$ speaker groups for the word $\mathbf{W}$ be given by $\Lambda_{L1}(\mathbf{W}_q)$ and $\Lambda_{L2}(\mathbf{W}_q)$, respectively. Using Eq. (10), $\Lambda_{L1}(\mathbf{W}_q)$ and $\Lambda_{L2}(\mathbf{W}_q)$ are given by:

$$\begin{aligned}\log(\Lambda_{L1}(\mathbf{W}_q)) &= \sum_{i=1}^{N+1} \log\left(p_{\mathbf{w}}^{L1}(O_{qi}|S_{qi})p_{\mathbf{w}}^{L1}(S_{qi}|S_{qi-1})\right),\\ &= \sum_{i=1}^{N+1} \log\left(\Lambda_{L1}(S_{qi},O_{qi})\right)\end{aligned} \tag{12}$$

and

$$\begin{aligned}\log\left(\Lambda_{L2}(\mathbf{W}_q)\right) &= \sum_{i=1}^{N+1} \log\left(p_{\mathbf{w}}^{L2}(O_{qi}|S_{qi})p_{\mathbf{w}}^{L2}(S_{qi}|S_{qi-1})\right),\\ &= \sum_{i=1}^{N+1} \log(\Lambda_{L2}(S_{qi},O_{qi})),\end{aligned} \tag{13}$$

respectively. In order to develop the differential model, we focus on the ith single element of the Markov chain as shown in Fig. 4. From Eqs. (12) and (13), it is seen that the contributions of the ith element to the overall $L1$ and $L2$ log-likelihoods are given by $\log(\Lambda_{L1}(S_{qi},O_{qi}))$ and $\log(\Lambda_{L2}(S_{qi},O_{qi}))$. For a single element (S_{qi},O_{qi}), we define the normalized delta log-likelihood as:

$$\Lambda_{\delta}(S_{qi},O_{qi}) = \frac{\log(\Lambda_{L1}(S_{qi},O_{qi})+1) - \log(\Lambda_{L2}(S_{qi},O_{qi})+1)}{\log(\Lambda_{L1}(S_{qi},O_{qi})+1) + \log(\Lambda_{L2}(S_{qi},O_{qi})+1)}, \tag{14}$$

where $-1 \leqslant \Lambda_{\delta}((S_i,O_i)) \leqslant +1$. From Eq. (14), a value of $\Lambda_{\delta}((S_i,O_i)) \rightarrow -1$ indicates an articulation leaning towards L2. Alternatively, $\Lambda_{\delta}((S_i,O_i)) \rightarrow +1$ indicates a more L1 like articulation. The delta likelihood score for the entire word $\mathbf{W}$ can be conveniently expressed as an average of the individual state-occupancy delta likelihoods,

$$\Lambda_{\delta}(\mathbf{W}_q) = \frac{1}{N}\sum_{i=1}^{N}\Lambda_{\delta}(S_{qi},O_{qi}), \tag{15}$$

where N is the total number of states. Clearly, $-1 \leqslant \Lambda_{\delta}(\mathbf{W}_q) \leqslant +1$. Since the term $\Lambda_{\delta}(\mathbf{W}_q)$ is bounded, it allows for a straight-forward articulation comparison across words and speakers. Specifically, by fixing a word the pronunciation of various individuals among $L1$ and $L2$ groups can be ordered on the scale $[-1,+1]$ which serves as a measure of accentedness. It is noted that $\Lambda_{\delta}(\mathbf{W}_q)$ in Eq. (15) is the accentedness score for $\mathbf{W}$ and the qth HF-type. The overall accentedness $\Lambda_{\delta}(\mathbf{W})$ can be computed by averaging $\Lambda_{\delta}(\mathbf{W}_q)$ across all HF-types,

$$\Lambda_{\delta}(\mathbf{W}) = \frac{1}{M}\sum_{q=1}^{M}\Lambda_{\delta}(\mathbf{W}_q), \tag{16}$$

where M is the total number of HF-types.

The proposed Markov process based accent model is adept at capturing the major articulatory differences between linguistically disparate population groups as well as specific articulation traits of individuals. As a result, the proposed scheme forms an extremely useful tool for individuals trying to acquire a new language or neutralize their accent. Additionally, the proposed scheme also serves as an excellent scientific tool for rapid automatic accent analysis of population groups.

5. Results and discussion

The experimental evaluations presented in this section use the data from N-AE (native American English) and N-MC (native Mandarin Chinese) speakers in the CU-Accent corpus. To facilitate a cross-sectional study, the N-MC speakers are divided into two groups: N-MC 1 and N-MC 2 based on their L2-exposure. Particularly, the N-MC 1 and N-MC 2 groups have L2-exposures of less than and greater than 2 yr. Furthermore, the 3 speaker groups are also divided into test and train groups. The train and test groups for N-MC speakers consisted of 19 and 18 distinct speakers, respectively. Similarly, the train and test groups for N-AE speakers consisted of 56 and 4 distinct individuals, respectively. In this manner, the accent analysis system is trained on 75 and tested on 22 separate speakers. In other words, the train and test sets consist of 8740 and 2530 samples of isolated word utterances. Additionally, the above mentioned test speaker set (18 N-MC + 4 N-AE speakers) is also used for both human as well as machine evaluation. In this manner, the automatic algorithm and human listeners evaluate the same speakers which allows for a direct comparison and assessment of the proposed system. Finally, the data from the test group speakers (AE and MC) was independent of the training data that was used for building the HF classifier as well as the HF-based articulation models.

5.1. Listener evaluation

In this experiment, 10 independent N-AE listeners were asked to rate the accents of speakers belonging to the above-mentioned speaker-groups. Data from the above-described test group (consisting of 22 speakers) were used for the listener evaluation, with 4, 9 and 9 speakers from the N-AE, N-MC 1 and N-MC 2 groups. A total of 103 speech tokens were presented to the listeners, and the listeners were asked to rate the accent level heard on a continuous scale of 0-to-100. Here, a score of 0 represents a heavy foreign accent, and 100 represents no-perceived accent (i.e., native AE speaker), respectively. The 103 tokens consisted of data from N-MC 1, N-MC 2 and N-AE speakers presented in a randomized fashion. The listener test duration was designed to last less than 20 min in order to minimize listener fatigue in the evaluation. Hence, the listening material was limited to use only 4 words for each speaker. Here, it is also noted that it is important to use the same words across all speakers since the goal of this study is to develop a system that can compare the accents of individuals. The use of different words for different speakers can potentially introduce biases in the judgement of both humans and the proposed algorithm. Furthermore, each token used in the listener evaluation was built by concatenating isolated utterances of 1-of-4 words,namely,target,three,thirty,and hear. Particularly, 3 separate utterances of the same word by the same speaker were concatenated to form each token. For example, 3 repetitions of the word thirty were combined with pauses in-between to form a token (i.e., thirty-pause-thirty-pause-thirty). This concatenation was performed in order to increase the speech material presented to the human listeners. Finally, the human-assigned accentedness score for each speaker was computed by taking an average across all their token-scores. The listener test design used for our study is very similar to that previously used by Flege (1988). It is also noted that the choice of words in the listener evaluation was arbitrary since previous studies on the CU-Accent corpus had shown that listeners are capable of detecting accents across all isolated word utterances within the corpus (Arslan and Hansen, 1996a). Additionally, the design of the listening study ensures that humans and machine evaluate the same set of speakers. However, while the humans heard only a subset of words per speaker (4 out of 23) for judging the accent, the machine processed all 23 words.

In Fig. 5(a), the relation between average human-assigned accentedness scores and L2-exposure of N-MC speakers is shown. For completeness, the average human-assigned accent scores of N-AE speakers are also included in the figure. Additionally, the average accentedness scores for each speaker-group is also shown. From the figure, it is observed that the average human-assigned accentedness scores for N-MC 1, N-MC 2, and N-AE groups are 52, 63, and 97, respectively. The numbers exhibit an increasing trend where the accent proficiency of non-native speakers increases with increased L2-exposure. This corroborates well with earlier research which has shown that perceived accent tends to reduce with increased L2-exposure (Flege, 1988). An overall correlation of 0.2 is observed between N-MC speaker scores and L2-exposure.

On the lines of human-assigned scores, Fig. 5(b) shows the relationship between machine-assigned accent scores and L2-exposure of N-MC speakers. The machine-assigned scores are computed for each speaker by averaging the normalized delta-likelihoods (using Eq. (16)) across all HF-dimensions and corpus-vocabulary (23 words). In Fig. 5 (b), the average Λ_δ scores for each speaker group is also shown. As seen in the human-assigned scores, the machine-assigned scores also exhibit a trend of increasing proficiency with greater L2-exposure with a correlation of 0.39 ($p < 0.05$) between N-MC speakers and L2-exposure.

Finally, the correlation between machine-assigned and human-assigned accentedness scores for both N-AE and N-MC speakers is determined to be 0.8 with high statistically significance ($p < 0.0001$). This result suggests a very high degree of agreement between human and machine assigned scores. Additionally, Fig. 5(c) also shows the scatter plot of machine-assigned *vs.* human-assigned scores. The data-points of the three speaker groups are shown separately in the scatter-plot (native AE (N-AE), and low L2 exposure < 2 yr N-MC-Group 1, and > 2 yr as N-MC-Group 2). The plot generally indicates a linear relationship between human-assigned and machine-assigned accentedness scores.

5.2. Differential analysis

As the proposed accentedness model is based on transitional as well as durational properties of articulation, it is possible to study each independently. In this experiment, the accentedness scores are recomputed separately using only the transitional and durational aspects of the proposed accent model. The transition-only and duration-only accentedness scores for all test-speakers are shown in Figs. 6 and 7. Furthermore, as seen in the figures the accentedness scores are also computed along each HF-dimensions separately, using Eq. (15). Particularly, the figures shown are for the scores along 'degree', 'height', 'frontness', and 'place' HFs. Finally, the correlation of the accentedness scores with the L2-exposure of N-MC speakers is also shown in the figures.

A general comparison of the correlation coefficients for durational and transitional aspects of articulation shows that *durational proficiency* is acquired faster than *transitional proficiency*. The observation can be made from 'degree' HF (Figs. 6(b) and 7(b)) where it is observed that the durational-correlation (0.36) is much higher than transitional-correlation (0.10). It may be useful to recall that since the 'degree' HF property encompasses all phone-types (vowel, closure, fricative, approximant, and flaps), it provides a broad overview of the durational and transitional aspects of speech. Therefore, it suggests that the proficiency in the durational aspect of vowel, fricative, approximant, closure and flap has increased more rapidly than the transitional aspects. This observation seems to corroborate with the findings of a related study, where temporal patterns formed the dominant cue in identifying AE vowels among N-MC individuals (Flege et al., 1997). Furthermore, the importance of temporal cues was seen to diminish in importance with increasing L2-exposure. As N-MC speakers perceive temporal differences early on, the durational aspects of articulation can be expected to acquire native-like characteristics rather quickly with increasing L2-exposure.

It is also observed that among consonants, the 'place' HF shows a higher durational-correlation (0.25, in Fig. 7(a)) than transitional-correlation (0.04, in Fig. 6(a)). From Fig. 6(c) and Fig. 7(c), it is also seen that the 'height' (vertical-position-of-tongue) HF shows the largest improvements in durational and transitional proficiency with increased L2-exposure (correlation coefficients: 0.61 for transition and 0.37 for duration). Alternatively, a

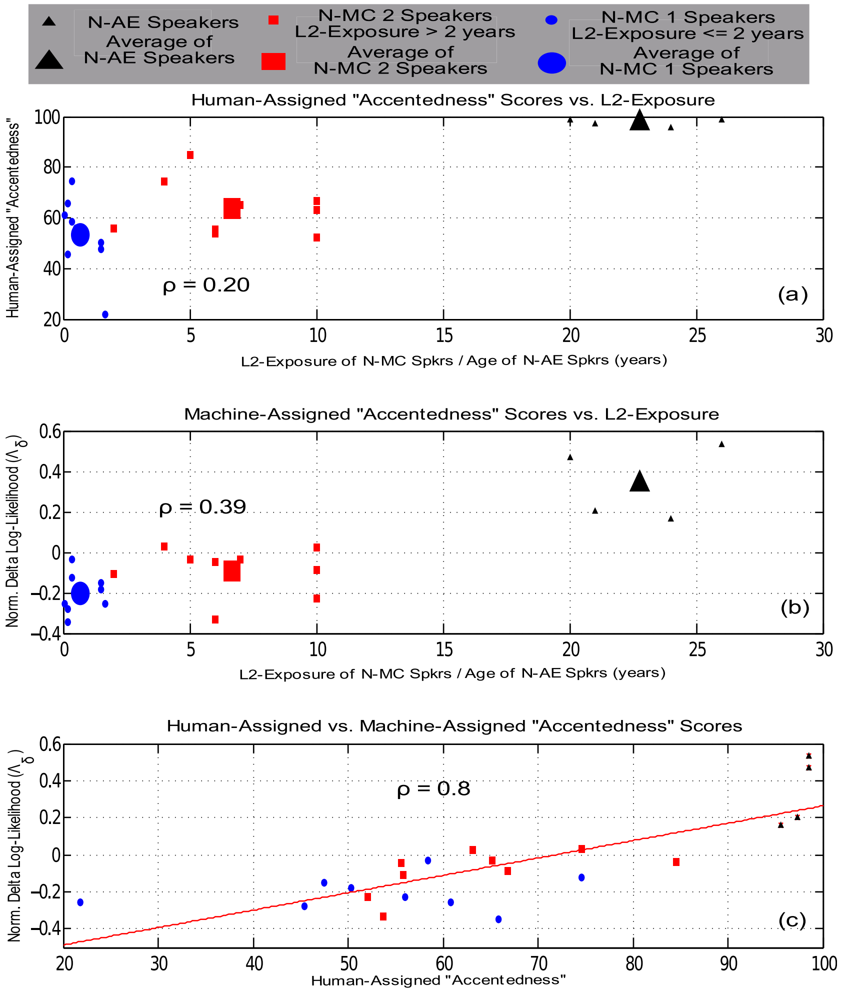

Fig. 5. Illustrates the relationship between (a) Human-assigned accentedness scores *vs.* L2-exposure of N-MC speakers (only N-MC speakers used to compute correlation coefficient), (b) machine-assigned accentedness scores (only N-MC speakers used to compute correlation coefficient) *vs.* L2-exposure of N-MC speakers, and (c) human-assigned *vs.* Machine-assigned accentedness scores (both N-AE and N-MC speakers used to compute correlation coefficient). Both human and machine assigned accentedness scores show increasing proficiency among N-MC speakers with greater L2-exposure. The human and machine assigned scores exhibit a large correlation (0.8) which is statistically significant ($p < 0.0001$).

relatively lower increase in proficiency is observed for ‘frontness’ (horizontal-position-of-tongue) HF.

While Figs. 6 and 7 highlight the overall proficiency picture of N-MC speakers, it is more interesting to present specific articulatory challenges that distinguish non-native from native speakers. Here, every word to be pronounced is viewed as an articulatory challenge, as it requires the speaker to move through a series of canonical articulatory states while maintaining the necessary durational and transitional constraints. Here, the required articulation due to canonical requirements is termed as the target articulation. The non-native speakers pronunciation of the canonical requirements is termed as the produced articulation. The higher the gap between target and produced articulation, the more difficult the canonical requirements are for the non-native speakers. The gap between target and produced articulation is effectively measured by using the proposed accent models and Eq. (16). First, the normalized delta-log-likelihood scores are collected for all target-produced articulation pairs, and the average score for each pair is computed. Subsequently, the pairs exhibiting the largest Λ_δ scores are collected as the most challenging for non-native speakers. While large Λ_δ scores indicate the severity of the mismatch, it is also important to measure the relative frequency of occurrence of the mismatch, *i.e.*, the number of times non-native speakers apply a certain produced-articulation when faced with a specific target-articulation. For this purpose, the relative frequency of occurrence of produced-target articulation pairs is measured. Here, the relative frequency of occurrence for a target-produced articulation-pair is defined as the ratio of the number of times a target-produced pair is observed when a target-articulation was required to be articulated.

The results of the above-described experiment are presented in Table 4. Table 4 shows the most difficult articulatory challenges (column: target articulation) faced by N-MC speakers in terms of ‘degree’, ‘place’, ‘height’, and ‘frontness’ HF-types. The table also shows the produced articulation for each target articulation along with the corpus words that contain the target articulation. It may be noted that the part of the word that constitutes the target articulation is underlined for each example corpus word shown. Finally, the average Λ_δ scores accumulated by N-MC-1 and N-MC-2 speakers for a particular target articulation-produced articulation pair is also shown. Finally, the relative frequency of occurrence (expressed as a percentage) for each target-produced articulation pair and both N-MC-1 and N-MC-2 speakers is also presented. In what follows, the results in Table 4 are discussed in detail.

For the case of 'degree' HF, it is seen that word-beginning and word-ending approximants and closures present articulatory difficulties for N-MC speakers (M.1, M.2, M.3, M.4, and M.7). These difficulties are observed in the N-MC pronunciations of 'root', 'white', 'look', 'like', 'hear', and 'there'. Here, the semi-vowels /l/, /w/, and /r/ (identified by underlined sections of the words) seem to challenge N-MC speakers where their pronunciations exhibit articulatory state substitutions. For example, in M.1 and M.2 the N-MC speakers are faced with the same articulation requirement, namely, silence approximant vowel, and are observed to substitute approximant with closure or fricative to produce silence fricative vowel (67% of times) or silence closure vowel (23% of times). It is also interesting to note that among all the above-mentioned articulatory challenges, an improvement in Λ_δ scores is observed when comparing N-MC 1 to N-MC 2 speakers. Additionally, M.5 shown in Table 4 presents articulation that involves word-ending stop-consonants. Here, N-MC speakers are observed to substitute closure in vowel closure fricative with silence to produce vowel silence fricative. Previous studies have reported that N-MC speakers exhibit longer closure durations in word-final stop-consonants (Arslan and Hansen, 1996b). The reason attributed to this observation was the absence of word-final voiced/unvoiced consonants in Mandarin. In our analysis, a similar observation is made from M.5, where the closure part of word-final stop-consonant is detected as silence (as longer VOT times are confused to be silence by the HF extraction system).

For the 'place' HF, it is observed from Table 4 that the 'rhotic' and 'lateral' place-of-articulations pose the greatest articulatory challenge for N-MC speakers. In the corpus vocabulary, 'rhotic' occurs in 'hear' and 'there'; and 'lateral' occurs in 'line' and 'look'. Additionally, 'place' HF

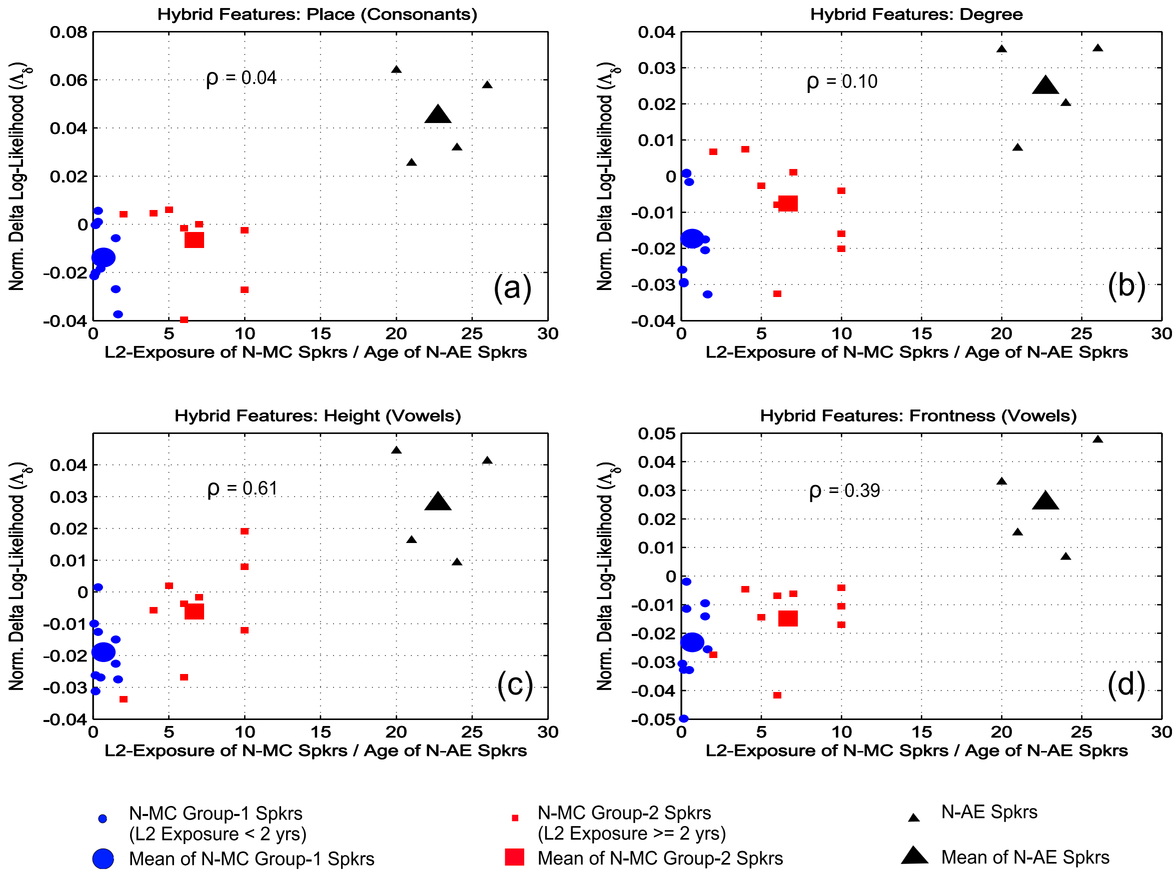

Fig. 6. The relationship between machine-assigned accentedness scores *vs.* L2-exposure of N-MC speakers when only the transitional aspect of articulation is considered along with the following HF-types: (a) place, (b) degree, (c) height, and (d) frontness. Transitional aspects of vowel articulation indicates a trend where greater improvement in proficiency as compared to consonants is observed for N-MC speakers (only N-MC speakers used to compute correlation coefficients).

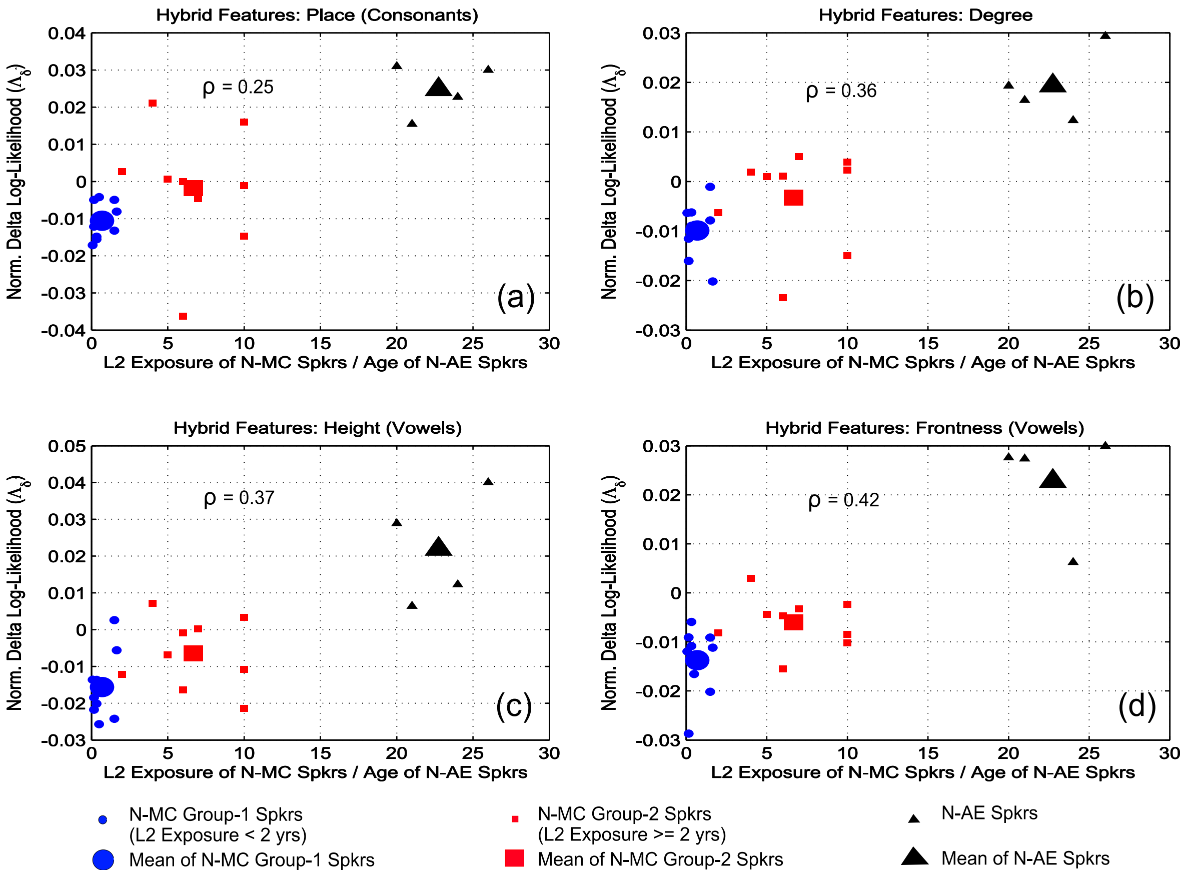

Fig. 7. The relationship between machine-assigned accentedness scores *vs.* L2-exposure of N-MC speakers when only the durational aspect of articulation is considered along with the following HF-types: (a) place, (b) degree, (c) height, and (d) frontness. Durational aspects of articulation indicates a trend where greater improvement in proficiency as compared to transitional aspects is observed for N-MC speakers (only N-MC speakers used to compute correlation coefficients).

shows very marginal improvement in proficiency as compared to 'degree' HF. This observation can be made by comparing the Λ_δ scores of N-MC 1 and 2 speaker groups for 'place' and 'degree' HFs. Additionally, the rhotic place-of-articulation in none rhotic silence is substituted by dental/alveolar 79% of times, and lateral in silence lateral none is substituted by labio-dental/labial 88% of times.

In terms of vowels, the greatest difficulty in articulation emerges in /iy/ *vs.* /ih/ (*e.g.*, beet *vs.* bit). This observation can be made from H.1 and F.1 in Table 4 (thirty and three). It is observed that the tongue height moves from a very-high to high position (in H.1), and tongue frontness moves from front to mid-front position (in F.1) with more than 90% frequency of occurrence. It is interesting to note that several researchers have highlighted this confusion (/iy/ *vs.* /ih/) for N-MC speakers in the past (Flege et al., 1997).

In addition to the articulatory challenges related to state-transitions, the articulatory durational challenges are presented below. It is expected that the average duration spent in any articulatory state would differ for native and non-native speakers. Such a difference in duration would emerge as a difference in durational Λ_δ scores when the average native and non-native speaker scores are compared. Therefore, it is inferred that the articulatory states which exhibit the largest difference in durational Λ_δ scores would pose the biggest durational challenge to non-native speakers. Fig. 8 shows the average durational Λ_δ scores of N-MC-1, N-MC-2, and N-AE speaker groups across the entire corpus. Particularly, Fig. 8(a) shows the Λ_δ scores of the 3 speaker-groups for 'place' HF. Here, it is observed that the durational aspect of 'lateral' place-of-articulation forms the biggest articulatory challenge for N-MC speakers (biggest gap in N-AE and N-MC scores). On the other hand, the 'velar' place-of-articulation shows the largest increase in durational proficiency as opposed to the other places-of-articulation (N-MC-2 is more closer to N-AE than N-MC-1). Moving on, the average 'height' HF Λ_δ scores for all 3 speaker groups are shown in Fig. 8(b). Here, durational aspects of 'low' and 'mid-low' tongue-positions pose a significant challenge. These two tongue-positions also exhibit the greatest proficiency improvements when comparing N-MC-1 with N-MC-2 speakers. On the other, the durational aspects of 'mid' tongue-position are found to be a lesser articulatory challenge. Fig. 8(c) shows the comparison of average Λ_δ scores for the 3 groups for 'frontness' HF. Here, the extreme tongue-positions of 'front' and 'back' pose a greater challenge than intermediate tongue-positions. Finally, the 'degree' HF in

Table 4
Articulatory transitions in American English that pose a greater difficulty to Native speakers of Mandarin Chinese. Larger average N-MC-1 and N-MC-2 scores indicate an increased difficulty level posed by the target articulation. Large differences in N-MC-1 (L2-exposure < 2 yr) and N-MC-2 (L2-exposure > 2 yr) scores indicate greater improvements in proficiency with increased L2-exposure. A large frequency of occurrence indicates a larger number of N-MC speakers substitute target-articulation with produced articulation.

HF-type		Target articulation	Produced articulation	Corpus words with target articulation	Average N-MC 1	Average N-MC 2	Occurrence frequency
Degree	M.1	SIL Approximant Vowel	SIL Fricative Vowel	root, white	−0.151	−0.111	67%
	M.2	SIL Approximant Vowel	SIL Closure Vowel	root, white	−0.071	−0.050	23%
	M.3	SIL Closure Vowel	SIL Fricative Vowel	look, line	−0.135	−0.100	63%
	M.4	SIL Closure Vowel	SIL Approximant Vowel	look, line	−0.038	−0.035	16%
	M.5	Vowel Closure Fricative	Vowel SIL Fricative	root, white, would, communication, look, feet, pump, change, target	−0.069	−0.063	97%
	M.6	Fricative Approximant Closure	Fricative Vowel Closure	thirty, bird	−0.057	−0.042	100%
	M.7	Vowel Approximant SIL	Vowel Closure SIL	hear, there	−0.037	−0.033	73%
Place	P.1	None Rhotic SIL	None Dental SIL	hear, there	−0.126	−.0120	11%
	P.2	None Rhotic SIL	None Alveolar SIL	hear, there	−0.042	−0.041	68%
	P.3	SIL Lateral None	SIL Labio-Dental None	line, look	−0.050	−0.045	81%
	P.4	SIL Lateral None	SIL Labial None	line, look	−0.041	−0.032	7%
Height	H.1	None Very-High SIL	None High SIL	thirty, three	−0.096	−0.089	93%
Frontness	F.1	None Front SIL	None Mid-Front SIL	thirty, three	−0.030	−0.025	92%

Fig. 8(d) clearly shows that the durational aspects of ‘approximants’ are the hardest to articulate followed by ‘vowels’. On the other hand, durational requirements of ‘fricatives’ and ‘closures’ are easily met by N-MC speakers. ‘Approximants’ also constitute the greatest improvements in proficiency followed by ‘vowels’.

The detailed analysis presented here serves to illustrate the power of the proposed accent analysis technique. It highlights the capability of the proposed system to automatically extract a wide range of useful articulation information. The information provided highlights the articulatory challenges that pose greater difficulty to non-native speakers both in terms of transitional and durational requirements. This information is provided as a higher gross level picture in Figs. 5–7. Additionally, the same information is provided in far greater details in Tables 4 and Fig. 8. Interestingly, several findings regarding N-MC accented AE in this study corroborate well with previous research. The strength of the proposed technique is the capability to automatically and efficiently extract these patterns. In this manner, the proposed accent analysis system is highly beneficial for language learners, speech scientists as well as engineers working in speech technology.

6. Conclusion

In this study, an accent analysis system based on PFs (phonological features) was presented. The use of PFs as a framework for analysis was strongly motivated by the capability of PFs to capture the fine articulatory variations in speech. It was argued that since the origins of accent are strongly embedded in speech production, PFs would form an ideal platform for analysis. In the study presented, two aspects of articulation were exploited to model accent, namely, articulatory transitions and articulatory durations. Both the transitional and durational aspects of articulation were directly extracted from the PF sequences. Furthermore, a Markov model based scheme was proposed to model the transitional and durational aspects of accent. It was proposed that this Markov model formulation could be used for learning native as well as non-native articulation traits. Finally, the native and non-native articulation models were used to automatically generate an “accentedness” score. The proposed accentedness score was used to measure accent proficiency of a speaker on a scale of $[-1,+1]$ where -1 corresponds to highly non-native-like and $+1$ to native-like pronunciation. Using the basic definition of the proposed accentedness measure, a comprehensive accent analysis methodology was presented where both gross and fine levels of articulation differences between native and non-native speakers could be automatically extracted. In particular, the following aspects of articulation were studied in isolation: (i) transitional, (ii) durational, (iii) vowel, (iv) consonants, (v) tongue-height, (vii) tongue-frontness, (viii) place-of-articulation, and (ix) manner-of-articulation.

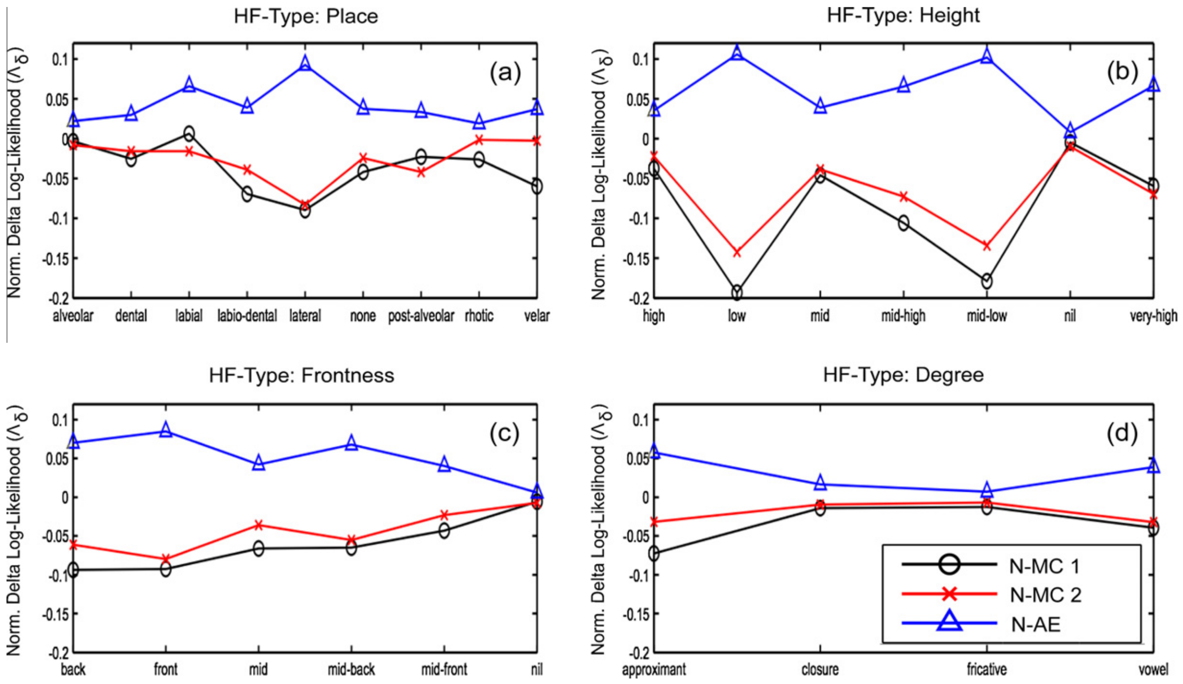

Fig. 8. Articulatory challenges faced by N-MC speakers in terms of duration requirements. The average Λ_δ scores of N-AE, N-MC 1, and N-MC 2 speaker groups for each HF-state is presented for the following HF-types: (a) place, (b) height, (c) frontness, and (d) degree. A larger gap in N-AE and N-MC scores corresponds to a greater articulatory difficulty. A larger gap in N-MC 1 and N-MC 2 scores reflects the largest improvement in proficiency with increase L2-exposure.

In this study, the proposed accent analysis system and methodology was applied in contrasting accent characteristics of N-MC (native speakers of Mandarin Chinese) and N-AE (native speakers of American English) speakers. The accentedness scores obtained automatically via the proposed system were compared to human-assigned counterparts. Strong agreement between the two set-of-scores was found (*i.e.*, correlation equal to 0.84 with high statisltical significance ($p < 0.0001$)). Furthermore, a gross level articulation analysis revealed that (i) N-MC proficiency increased with L2 exposure, (ii) durational aspects of articulation improved faster than transitional aspects with increased L2 exposure, and (iii) proficiency in vowel articulation improved and proficiency in consonant articulation stagnated with increasing L2 exposure. A finer level analysis revealed that (i) semi-vowels such as /r/, /l/, and /w/ pose articulatory challenges to N-MC speakers in terms of articulator motion and duration. Among vowels and diphthongs, /iy/ *vs.* /ih/ along with diphthongs /ay/ and /ey/ constitute major articulatory challenges to N-MC speakers. In terms of place-of-articulation, 'rhotic' and 'lateral' positions are most challenging. In general, word-final stop-consonants are difficult for N-MC speakers to articulate.

The proposed accent analysis system and methodology could potentially provide useful automatically derived strength and weakness feedback to students using language learning systems. Additionally, this system can also be used to study general articulatory characteristics of linguistics groups. It could be used to provide a rapid first-level analysis to speech scientists, thereby focussing their attention to critical articulator disparities.

For future work, the experiments presented in this paper could be expanded to include phrases and spontaneous speech elements of the CU-Accent corpus. The use of spontaneous speech should further enhance the capability of the proposed system. Further work towards incorporating a perceptual element in the proposed system is also of interest. Such an addition would be beneficial, as not all articulatory differences are equally relevant to human perception. Incorporating a perceptual element in the model should further improve the agreement between human-assigned and machine-generated accentedness scores.

Acknowledgement

This work was supported by the USAF under a subcontract to RADC, Inc., Contract FA8750-09-C-0067. (Approved for public release. Distribution unlimited.)

References

Angkititrakul, P., Hansen, J.H., 2006. Advances in phone-based modeling for automatic accent classification. IEEE Trans. Audio Speech Lang. Process. 14 (2), 634–646.

Arslan, L.M., Hansen, J.H., 1996a. Language accent classification in American English. Speech Comm. 18 (4), 353–367.

Arslan, L.M., Hansen, J.H., 1996b. A study of temporal features and frequency characteristics in American English foreign accent. J. Acoust. Soc. Amer. (JASA) 102, 28–40.

Choueiter, G., Zweig, G., Nguyen, P., 2008. An empirical study of automatic accent classification. In: ICASSP, pp. 4265–4268.

Chreist, F., 1964. Foreign Accent. Prentice-Hall, Englewood Cliffs, NJ.

Das, S., Hansen, J.H., 2004. Detection of voice onset time (VOT) for unvoiced stops (/p/,/t/,/k/) using the Teager energy operator (TEO) for automatic detection of accented English. In: IEEE NORSIG: Northern Symp. on Signal Processing, pp. 344–347.

Flege, J., 1988. Factors affecting degree of perceived foreign accent in English sentences. J. Acoust. Soc. Amer. (JASA) 84 (1), 70–79.

Flege, J., Bohn, O.-S., Jang, S., 1997. Effects of experience on non-native speakers production and perception of English vowels. J. Phonetics 25, 437–470.

Frankel, J., Magimai-Doss, M., King, S., Livescu, K., Cetin, O., 2007a. Articulatory feature classifiers trained on 2000 h of telephone speech. In: Interspeech.

Frankel, J., Wester, M., King, S., 2007b. Articulatory feature recognition using dynamic bayesian networks. Comput. Speech Lang. 21 (4), 620–640.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V., 1993. TIMIT acoustic-phonetic continuous speech corpus. LDC93S1, LDC. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1>.

Hansen, J.H.L., Gray, S., Kim, W., 2010. Automatic voice onset time detection for unvoiced stops (/p/, /t/, /k/) with application to accent classification. Speech Comm. 52 (10), 777–789.

Jia, G., Strange, W., Wu, Y., Collado, J., Guan, Q., 2006. Perception and production of English vowels by Mandarin speakers: age-related differences vary with amount of l2 exposure. J. Acoust. Soc. Amer. (JASA) 119 (2), 1118–1130.

Jou, S.-C., Schultz, T., Waibel, A., March 2005. Whispery speech recognition using adapted articulatory features. In: ICASSP, pp. 1009–1012.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Comput. Speech Lang. 14 (4), 333–353.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Amer. (JASA) 121 (2), 723–742.

Leung, K., Mak, M., Siu, M., Kung, S., 2006. Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. Speech Comm. 48 (1), 71–84.

Mak, B., Siu, M., Ng, M., Tam, Y., Chan, Y., Leung, K., Ho, S., Chong, F., Wong, J., Lo, J., 2003. Plaser: pronunciation learning via automatic speech recognition. In: Human Language Technology Conf., Vol. 2. pp. 217–220.

Mangayyagari, S., Islam, T., Sankar, R., 2008. Enhanced speaker recognition based on intra-modal fusion and accent modeling. In: Internat. Conf. on Pattern Recognition.

Markov, K., Dang, J., Nakamura, S., 2006. Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. Speech Comm. 48 (2), 161–175.

Metze, F., 2007. Discriminative speaker adaptation using articulatory features. Speech Comm. 49 (5), 348–360.

Metze, F., Waibel, A., 2002. A flexible streaming architecture for ASR using articulatory features. In: ICSLP.

Morris, J., Fosler-Lussier, E., 2008. Conditional random fields for integrating local discriminative classifiers. IEEE Trans. Audio Speech Lang. Process. 16 (3), 617–628.

Neri, A., Cucchiarini, C., Strik, H., 2006. ASR-based corrective feedback on pronunciations: does it really work ? In: Interspeech.

Pedersen, C., Diederich, J., 2007. Accent classification using support vector machines. In: 6th Internat. Conf. on Computer and Information Science.

Salvi, G., 2003. Using accent information in ASR models for Swedish. In: Eurospeech. pp. 2677–2680.

Sangwan, A., Hansen, J.H., 2007. Phonological feature based variable frame rate scheme for improved speech recognition. In: IEEE Automatic Speech Recognition and Understanding (ASRU), pp. 582–586.

Sangwan, A., Hansen, J.H., 2008. Evidence of coarticulation in a phonological feature detection system. In: Interspeech'08. pp. 1525–1528.

Scharenborg, O., Wan, V., Moore, R., 2007. Towards capturing fine phonetic variation in speech using articulatory features. Speech Comm. 49 (10-11), 811–826.

Tepperman, J., Narayanan, S., 2008. Using articulatory representations to detect segmental errors in nonnative pronunciation. IEEE Trans. Audio Speech Lang. Process. 16 (1), 8–22.

Wei, S., Liu, Q., Wang, R., 2006. Automatic Mandarin pronunciation scoring for native learners with dialect accent. In: Interspeech-06.

Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., Yoon, S.-Y., 2005. Accent detection and speech recognition for Shanghai-accented Mandarin. In: Interspeech-05. pp. 217–220.