INTERSPEECH 2009 BRIGHTON

On the use of Phonological Features for Automatic Accent Analysis

Abhijeet Sangwan and John H.L. Hansen

Center for Robust Speech Systems, University of Texas at Dallas

abhijeet.sangwan@utdallas.edu, john.hansen@utdallas.edu

Abstract

In this paper, we present an automatic accent analysis system that is based on phonological features (PFs). The proposed system exploits the knowledge of articulation embedded in phonology by rapidly build 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 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 rapidly perform quantitative as well as qualitative analysis of foreign accents. The work developed in this paper is easily assimilated into language learning systems, and has impact in the areas of speaker and speech recognition.

Index Terms: automatic accent analysis, phonological features

10.21437/Interspeech.2009-68

1. Introduction

Automatic accent analysis and classification is useful in many speech technologies such as automatic speech recognition (ASR), speaker recognition, and language learning [1]. Herein, building automatic accent classification systems has received a lot of attention where the use of cepstral features along with classifiers like GMMs (Gaussian mixture models), HMMs (hidden Markov models), and SVMs (support vector machines) is commonplace [3, 2]. Information helpful towards segregating accent-types has also been found in the temporal evolution of spectral features, and exploited appropriately in spectral trajectory models (STMs). Furthermore, low-level speech features such as VOT (voice-onset time), word/phone durations, intonation patterns, formant-behavior *etc.* have also been shown to assist is distinguishing between accents [3].

In this paper, we adopt a different approach towards modeling accent by capturing articulatory patterns of accented speech via phonological features (PFs). Specifically, we employ the hybrid features (HFs) proposed by Frankel *et. al.* owing to their direct relationship with articulatory phonetics [4]. The usage of PFs in opposition to cepstral (or spectral) features offers the advantage of assessing, comparing and contrasting articulatory aspects of different accents. Alternatively, given a word the articulation strategies inherent to different accents are explicitly revealed in the HF space.

In order to capture the realization of accent in HF space, we develop a Markov model (MM) that is based on transitional and durational characteristics of articulation. In this manner, different MMs are built for native and non-native speech by learning the articulatory strategies that are typically employed by native and non-native speakers, respectively. Subsequently, word articulations are assigned "accentedness" scores based on the likelihoods that they were produced by native or non-native models. As a result, the proposed PF-based system provides an easy methodology for conducting rapid compare and contrast analysis of different accents in terms of articulation. The proposed system could be a useful addition to CALL (computer aided language learning) where a learners articulation scores can be automatically generated for a large vocabulary and thereby used to track their progress. Furthermore, by systematically gathering major production characteristics of different accents the proposed system also allows for planning pronunciation modeling strategies for ASR systems. Finally, the accentedness scores generated by the proposed system could also be used as discriminatory input feature-set for speaker recognition systems.

In this paper, we demonstrate the performance of the proposed PF-based accent model in the task of assessing the degree of accentedness in English spoken by native speakers of Mandarin Chinese (N-MC) when compared to native speakers of American English (N-AE). This is accomplished by defining a new normalized likelihood measure based on the previously learnt MMs of native and non-native speech. The proposed normalized likelihood measure allows accentedness to be rated on a scale of -1 to $+1$, where a score tending to -1 and $+1$ indicate a highly non-native and native like pronunciation, respectively. Owing to the strong relation between production characteristics and perception of accent, the normalized likelihood measure is expected to have strong agreement with the degree of perceived accentedness. By dividing the N-MC speakers into two proficiency groups based on their exposure to American English (AE), we perform a cross-sectional study where the two N-MC groups are compared with N-AE speakers. As expected, the cross-sectional analysis of the N-MC groups reveal a general migration of accentedness scores toward native-like pronunciation with longer exposure of AE [6]. Furthermore, the use of HFs in our analysis model also reveals the major areas of changes from non-native to native like articulation in terms of vowel and consonant production characteristics. Finally, we also conduct an experiment where the proposed accentedness score across a fixed vocabulary of 22 words is used to predict the AE exposure of different N-MC speakers. Our experimental results show a reasonable correlation between estimated and actual exposures.

2. Proposed Accent Model

The PF dimensions within the HF system cover sufficient breadth and depth in phonology to simultaneously allow meaningful detection accuracy as well as analysis capability. Among the HFs: place, degree, nasality, rounding, and glottal serve mostly towards describing consonants characteristics; vowel,

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029

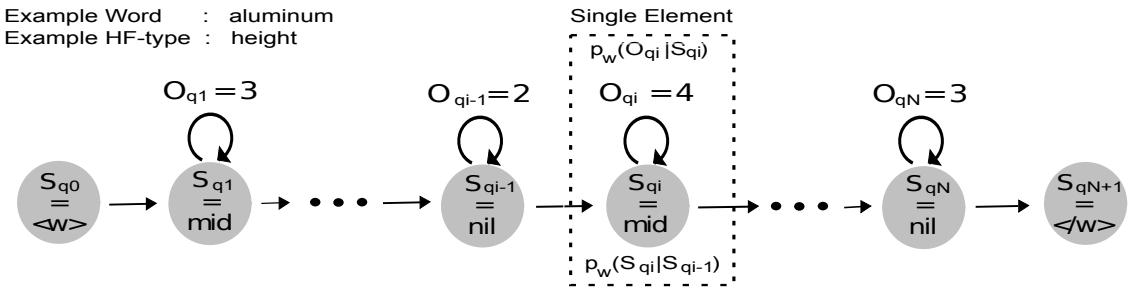

Figure 1: Proposed Markov model for learning transitional and durational aspects of accented articulation in HF (hybrid features) space.

height, and frontness deal exclusively with vowels. In this study, all dimensions of HFs are employed where the proposed MM is developed for each HF separately. In order to extract the HFs from speech utterances, we employ a HMM-based PF extraction scheme. The details of the HMM extraction scheme are given in [5]. It is useful to note that only utterances of native speakers were used to train the PF extraction scheme.

The HFs capture two important aspects of articulation, namely, the HF state-transitions as well as HF state-occupancy. For example, in the articulation of diphthong /aw/ in AE the tongue shifts from a low to high position while occupying each state for a finite duration of time. Herein, 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 HF-type value is 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 number of frames.

We propose the use of MM (Markov model) to model state-transitions and occupancies. The formulation of the MM is illustrated by the state-machines in Fig. 1. Since this study focuses on aspects of Mandarin accented English, we estimate the parameters of the Markov process separately for L1 (*i.e.*, English spoken by N-AE speakers) and L2 (*i.e.*, English spoken by N-MC speakers) groups. Herein, the L1 and L2 MMs establish extrema within the phonological (or articulation) space where individual articulation strategies can be gauged for their accentedness.

Articulating a word $\mathbf{W}$ involves production of a sequence of phones. In terms of HFs, articulation involves running 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
$\{(S_{q1}, O_{q1}), (S_{q2}, O_{q2}), ..., (S_{qN}, O_{qN})\}$
where the articulation is assumed to prolong over N states. The sequence of state-occupancy pairs forms a MM as shown in Fig. 1. In order to model the entry and exit into the MM, we define ordered pairs $(S_{q0}, 0)$ and $(S_{qN+1}, 0)$ as the entry and exit states. As shown in Fig. 1, the entry and exit states always carry the values $< w >$ and $< /w >$. 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:
$\{(S_{q0}, 0), (S_{q1}, O_{q1}), (S_{q2}, O_{q2}), ..., (S_{qN}, O_{qN}), (S_{qN+1}, 0)\}$.

Let the joint likelihood of observing the q^{th} HF states and occupancies be given by:

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

where $\mathbf{O}_q \equiv \{O_{q0}, O_{q1}, ..., O_{qN}, O_{qN+1}\}$ and $\mathbf{S}_q \equiv \{S_{q0}, S_{q1}, ..., S_{qN}, S_{qN+1}\}$ represent the series of occupancy and state observations respectively. Using conditional probability, Eq. (1) can be written as

$$\Lambda(\mathbf{W}_q) = p_{\mathbf{w}}(\mathbf{O}_q|\mathbf{S}_q)p_{\mathbf{w}}(\mathbf{S}_q) \qquad (2)$$

By assuming that the occupancy observations are mutually independent and only depend upon their respective states, the first part of Eq. (2) is given by:

$$p_{\mathbf{w}}(\mathbf{O}_q|\mathbf{S}_q) = \prod_{i=0}^{N+1} p_{\mathbf{w}}(O_{qi}|S_{qi}). \qquad (3)$$

Here $p_{\mathbf{w}}(O_{qi}|S_{qi})$ is the probability of observing an occupancy of O_{qi} frames given the state occupied is S_{qi}, and the word in question $\mathbf{W}$. It may be noted that where $p_{\mathbf{w}}(O_{q0}|S_{q0}) = p_{\mathbf{w}}(O_{qN+1}|S_{qN+1}) = 1$. 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. (2) is given by:

$$p_{\mathbf{w}}(\mathbf{S}_q) = \prod_{i=1}^{N+1} p_{\mathbf{w}}(S_{qi}|S_{qi-1}) \qquad (4)$$

where $p_{\mathbf{w}}(S_{qi}|S_{qi-1})$ is the transition probability due to Markov process assumption. By combining Eqs (3) and (4), we get the following expression for the joint likelihood in Eq. (1):

$$\Lambda(\mathbf{W}_q) = \prod_{i=1}^{N+1} p_{\mathbf{w}}(O_{qi}|S_{qi})p_{\mathbf{w}}(S_{qi}|S_{qi-1}). \qquad (5)$$

Alternatively, the log-likelihood of jointly observing the q^{th} HF states and occupancies are given by:

$$\log(\Lambda(\mathbf{W}_q)) = \sum_{i=1}^{N+1} \log(p_{\mathbf{w}}(O_{qi}|S_{qi})p_{\mathbf{w}}(S_{qi}|S_{qi-1})). \qquad (6)$$

In our study, $p_{\mathbf{w}}(O_{qi}|S_{qi})$ is assumed to be gamma distributed. We use the ML (maximum likelihood) estimates of the gamma-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

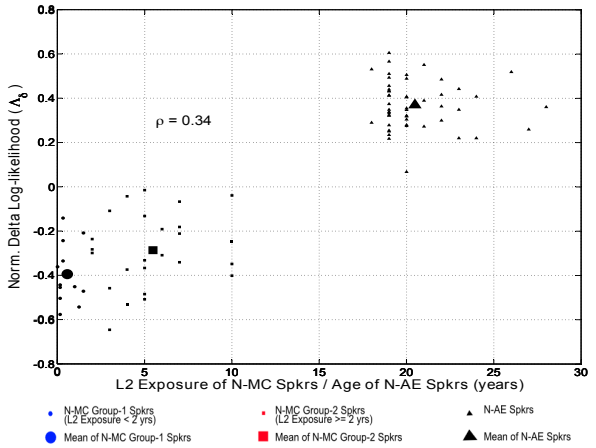

Figure 2: Normalized delta log-likelihood (Λ_δ) scores of N-MC and N-AE speakers *vs.* years of exposure to AE.

Table 1: HF-state transition: N-MC and N-AE proficiency

	Norm. Delta Log-Likelihood			
HF-type	N-MC 1	N-MC 2	N-AE	Correlation (ρ)
Place	-0.045	-0.037	0.042	0.25
Degree	-0.025	-0.018	0.025	0.2
Height	-0.035	-0.025	0.027	0.33
Frontness	-0.038	-0.024	0.030	0.34

Table 2: HF-state occupancy: N-MC and N-AE proficiency

	Norm. Delta Log-Likelihood			
HF-type	N-MC 1	N-MC 2	N-AE	Correlation (ρ)
Place	-0.022	-0.015	0.025	0.37
Degree	-0.015	-0.009	0.017	0.42
Height	-0.023	-0.017	0.022	0.25
Frontness	-0.021	-0.015	0.023	0.35

$p_{\mathbf{w}}(S_{qi}|S_{qi-1})$ are V and V^2, where V is the number of values the q^{th} HF-state can take. 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 uses $L1$ and $L2$ accent models.

Let the MMs for word $\mathbf{W}$ be given by $\Lambda_{L1}(\mathbf{W}_q)$ and $\Lambda_{L2}(\mathbf{W}_q)$ for $L1$ and $L2$ speaker groups, respectively. Using 6, $\Lambda_{L1}(\mathbf{W}_q)$ and $\Lambda_{L2}(\mathbf{W}_q)$ are given by:

$$\begin{aligned}\log(\Lambda_{L1}(\mathbf{W}_q)) &= \sum_{i=1}^{N+1} \log(p^{L1}_{\mathbf{w}}(O_{qi}|S_{qi})p^{L1}_{\mathbf{w}}(S_{qi}|S_{qi-1})),\\ &= \sum_{i=1}^{N+1} \log(\Lambda_{L1}(S_{qi},O_{qi})) \qquad (7)\end{aligned}$$

and

$$\begin{aligned}\log(\Lambda_{L2}(\mathbf{W}_q)) &= \sum_{i=1}^{N+1} \log(p^{L2}_{\mathbf{w}}(O_{qi}|S_{qi})p^{L2}_{\mathbf{w}}(S_{qi}|S_{qi-1})),\\ &= \sum_{i=1}^{N+1} \log(\Lambda_{L2}(S_{qi},O_{qi})) \qquad (8)\end{aligned}$$

respectively. In order to develop the differential model, we focus on the i^{th} single element of the Markov chain as shown in Fig. 1. From Eqs. (7) and (8), it is seen that the contributions of the i^{th} 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:

$$\Lambda_\delta(S_{qi},O_{qi}) = \frac{\log(\Lambda_{L1}(S_{qi},O_{qi})) - \log(\Lambda_{L2}(S_{qi},O_{qi}))}{\log(\Lambda_{L1}(S_{qi},O_{qi})) + \log(\Lambda_{L2}(S_{qi},O_{qi}))} \qquad (9)$$

where $-1 \leq \Lambda_\delta((S_i,O_i)) \leq +1$. From Eq. (9), a value of $\Lambda_\delta((S_i,O_i)) \rightarrow -1$ indicates an articulation leaning towards L2. On the other hand, $\Lambda_\delta((S_i,O_i)) \rightarrow +1$ 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, *i.e.*,

$$\Lambda_\delta(\mathbf{W}_q) = \mathbf{E}[\Lambda_\delta(S_{qi},O_{qi})] \qquad (10)$$

where $\mathbf{E}[.]$ is the expectation. Clearly, $-1 \leq \Lambda_\delta(\mathbf{W}_q) \leq +1$. Since the term $\Lambda_\delta(\mathbf{W}_q)$ is bounded, it allows straight-forward articulation comparisons 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.

3. Results and Discussion

In this paper, we compare the articulation characteristics of N-MC and N-AE speakers in the CU-Accent corpus [1]. We have divided the native-Mandarin Chinese (N-MC) speakers into two groups: N-MC 1 and N-MC 2 based on the number of years of residence in the U.S.A. While the N-MC 1 group speakers have been living in the U.S.A for less than 2 years, the N-MC 2 consists of speakers that have been living in the U.S.A for greater than 2 years. It is hypothesized that N-MC 1 would have lower English proficiency than N-MC 2 since N-MC 1 have had lower immersion time in the native AE accent [6]. All of our analysis in this paper is based on the cross-sectional study of 3 groups: N-MC 1, N-MC 2, and N-AE speakers,

We analyze the accent phenomenon within the HF space by comparing and contrasting the performance of the 3 study groups using the normalized delta log-likelihood (Λ_δ) scale proposed in Eq. (9). The use of Λ_δ as a statistical measure of accentedness allows systematic comparisons to be made (i) across the corpus, (ii) specific words/phrases, (iii) all HF-dimensions, and (iv) specific HF-dimensions (like place, height *etc.*). It is easy to see that the articulation characteristics could also be compared across various combinations of the above-enumerated conditions. Therefore, the proposed system allows for a comprehensive analysis in a drill-down fashion where the same measure (Λ_δ) can be studied in different context (general or specific) to reveal various details of articulation or accentedness. In Fig. 2, the normalized delta log likelihood (Λ_δ) scores for the 3 study groups across all HF dimensions and corpus-vocabulary are shown. Furthermore, the average Λ_δ scores for each study group is also shown. The trend of acquiring native like pronunciation characteristics with longer exposure to AE is easily seen as the average N-MC 2 score is higher than N-MC 1 score. Finally, the correlation coefficient (ρ) of N-MC speaker accentedness scores and their corresponding exposure to AE is computed to be 0.34 as indicated on Fig. 2. While Fig. 2 shows an overall picture of the study groups, Tabs. 1 and 2 show the transition and occupancy specific Λ_δ scores for 4 important HF-types: place, degree, height and frontness across each study group. Furthermore, the tables also show the

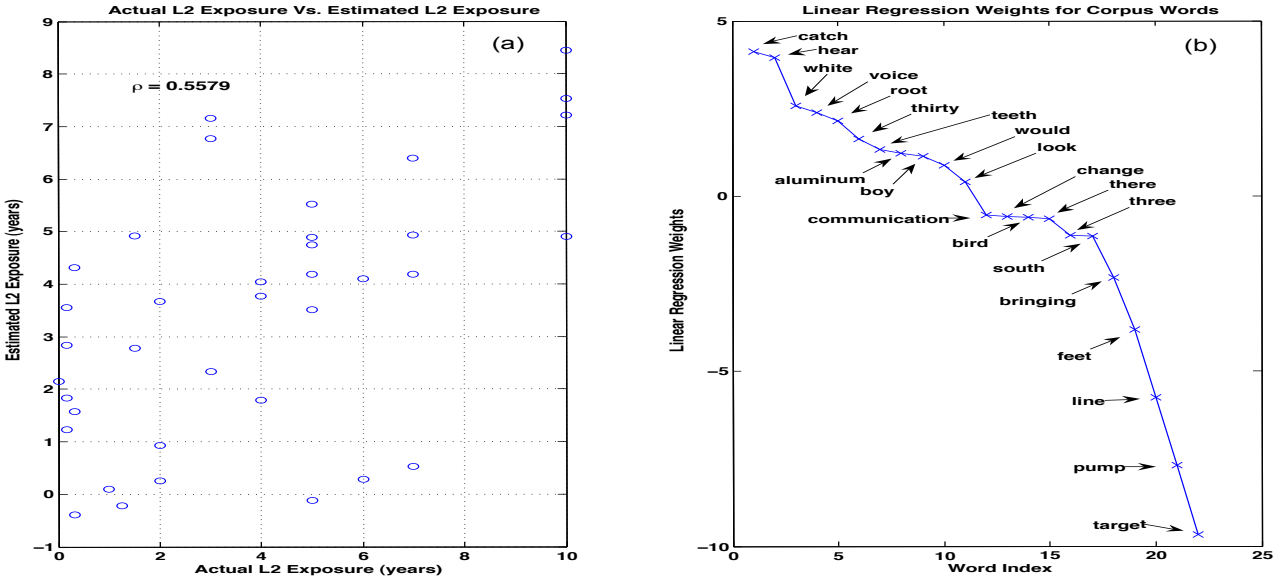

Figure 3: Estimating proficiency of N-MC speakers using linear prediction and Λ_δ scores generated by the proposed accent model: (a) Estimated *vs.* Actual AE exposure of N-MC speakers, and (b) linear prediction weights of 22 words used in predicting exposure.

correlation between AE exposure of N-MC speakers and the Λ_δ scores for each HF-type. In terms of HF-state transitions, it is observed from Tab. 1 that articulator transitions involving vowel production (height and frontness) have progressed more towards native-live production as compared to consonants (place). However, from Tab. 2, it can be seen that in terms of HF-state occupancy, vowels and consonants have migrated equally towards native-like production proficiency. Finally, by comparing the correlation coefficients for degree in Tabs. 1 and 2 a general trend is observed that indicates that N-MC speakers have acquired durational characteristics at a quicker rate than transitional characteristics.

In a final experiment, we have assessed the predictive power of the Λ_δ scores in estimating L2 exposure of N-MC speakers. Based on the above-presented analysis, the Λ_δ scores were available for 22 unique isolated word utterances in CU-Accent corpus per N-MC speaker. Using the Λ_δ scores for all 22 words per speaker as an input feature-set, a simple linear prediction (LP) model was trained to estimate the actual L2 exposures (available in corpus). Fig. 3 (a) shows the scatter-plot between the predicted AE exposure *vs.* the actual AE exposures of N-MC speakers. Also, the correlation coefficient of predicted *vs.* actual exposures is computed to be 0.558. Since exposure is related to accent, the above experiment also serves to indirectly demonstrate the quality of the proposed accentedness measure. A more direct assessment of the measure would be to observe the correlation between the machine-generated Λ_δ scores and human-generated accentedness scores in listener evaluations. Such a study is currently being conducted. In Fig. 3 (b), the LP weights of different words learnt by the prediction model are enumerated. The word weights reveal an interesting aspect, *i.e.*, words with strongly positive weights like catch, hear *etc.* show steady transition towards native-like pronunciation with increasing exposure. However, words with strongly negative weights like target, pump *etc.* continue to be difficult for N-MC speakers to pronounce despite increasing exposure to AE. Interestingly, in a accent classification task on the same corpus, the word 'target' was found to be particularly useful for correctly identifying MC speakers.

4. Conclusion

In this paper, we have developed and demonstrated a new approach towards automatic accent analysis. The proposed analysis methodology was based on the use of phonological features (PFs). The articulatory information in PFs was captured by means of a Markov model that learns the transitional and durational aspects of accented articulation. Furthermore, a new statistical measure of "accentedness" based on the Markov models of native and non-native articulation was also proposed. Finally, a cross-sectional study of accented English spoken by native speakers of Mandarin Chinese (N-MC) was performed to highlight the capability of the proposed system. Particularly, it was demonstrated that the proposed scheme is capable of (i) measuring English proficiency of N-MC speakers, (ii) identify significant areas of change and stagnation in articulation of non-native speakers with increasing immersion time, (iii) identify key-lexicon items that pose significant challenge for N-MC speakers, and (iv) predict exposure of a N-MC speakers based on their accentedness. The work presented in this paper is readily applied to language learning, and speaker recognition systems; and can be extended to assist in pronunciation modeling for ASR. The proposed system is also an attractive tool for speech scientists interested in rapid compare and contrast analysis of different accents.

5. References

[1] P. Angkititrakul and J. Hansen,"Advances in phone-based modeling for automatic accent classification", IEEE Trans. on Audio, Speech and Lang. Proc, vol. 14, 2006, pp. 634-646.

[2] C. Pedersen and J. Diederich, "Accent classification using support vector machines", 6^{th} Intl. Conf. on Comp. and Info. Sc., 2007.

[3] L.M. Arslan and John H.L. Hansen,"A study of temporal features and frequency characteristics in American English foreign accent", JASA, vol. 102, 1996, pp. 28-40.

[4] J. Frankel, M. Magimai-Doss, S. King, K. Livescu and O. Cetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech", Interspeech, 2007.

[5] A. Sangwan, A. Ikeno and J. Hansen,"Evidence of Coarticulation in a Phonological Feature Detection System", Interspeech, 2008.

[6] G. Jia, W. Strange, Y. Wu, J. Collado and Q. Guan,"Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure", JASA, vol. 119, 2006, pp. 1118-1130.