Automatic Accent Assessment Using Phonetic Mismatch and Human Perception

Freddy William, *Student Member, IEEE*, Abhijeet Sangwan, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—In this study, a new algorithm for automatic accent evaluation of native and non-native speakers is presented. The proposed system consists of two main steps: alignment and scoring. In the alignment step, the speech utterance is processed using a Weighted Finite State Transducer (WFST) based technique to automatically estimate the pronunciation mismatches (substitutions, deletions, and insertions). Subsequently, in the scoring step, two scoring systems which utilize the pronunciation mismatches from the alignment phase are proposed: (i) a WFST-scoring system to measure the degree of accentedness on a scale from -1 (non-native like) to $+1$ (native like), and a (ii) Maximum Entropy (ME) based technique to assign perceptually motivated scores to pronunciation mismatches. The accent scores provided from the WFST-scoring system as well as the ME scoring system are termed as the WFST and P-WFST (perceptual WFST) accent scores, respectively. The proposed systems are evaluated on American English (AE) spoken by native and non-native (native speakers of Mandarin-Chinese) speakers from the CU-Accent corpus. A listener evaluation of 50 Native American English (N-AE) was employed to assist in validating the performance of the proposed accent assessment systems. The proposed P-WFST algorithm shows higher and more consistent correlation with human evaluated accent scores, when compared to the Goodness Of Pronunciation (GOP) measure. The proposed solution for accent classification and assessment based on WFST and P-WFST scores show that an effective advancement is possible which correlates well with human perception.**

***Index Terms*—Automatic accent assessment, pronunciation scoring, finite state transducers (FST), maximum entropy models (MEMs), perception based measures.**

I. Introduction

EFFECTIVE pronunciation training for L2 learners can be delivered by training and intensive feedback which usually requires skilled and trained teachers [1]–[3]. However, this type of training is expensive and requires large amounts of time and commitment. In recent years, Computer Assisted Language Learning (CALL) and Computer Assisted Pronunciation Training (CAPT) applications which make use of Automatic Speech Recognition (ASR) have emerged as complementary tools that can automate proficiency assessment. In fact, CALL and CAPT applications can potentially automate assessment of a number of language skills such as pronunciation, fluency, and grammar. Reliable and automatic estimation of language specific speaking skills can be beneficial to language learners since it provides a flexible and customizable learning environment. This study focuses on developing a new automatic accent assessment algorithm which utilizes knowledge learned from human perception of accents to automatically score phone mismatches in pronunciation. In the past, a variety of accent measurements techniques have been developed, namely, Hidden Markov Model (HMM) log likelihood scores, segment classification error scores, segment duration scores, syllabic timing scores [4], [5], Goodness of Pronunciation (GOP) measure [6], linear and non-linear combination of confidence scores [7], and phonological features based pronunciation scores [8]. In two early studies [9], [10], an evaluation of speech features was conducted and employed for accent classification including voice-onset-time, vowel formant locations, phone duration, vowel spectral slope, stop release time, pitch slope, and other speech production features. HMM and GMM based classifiers were used for German, Turkish, and Chinese accents in American English. In [4], [5], four algorithms were explored. First, HMM log likelihood algorithm provides scores for non-native speech at phone level through log-likelihood scores for each phone segment, acquired from time alignment and Viterbi Algorithm using HMMs obtained from native speakers. Second, segment classification algorithm provides pronunciation score based on recognition accuracy of a phone, where the phone classifier is trained using native speakers. Third, a segment duration algorithm utilizes a measure of the rate of speech together with the duration of a segment to provide a score for non-native speech. Finally, syllabic timing scores are used to compute the normalized time between the center of vowels within segments of non-native speech to produce a syllabic timing score. Out of these four algorithms, the segment duration shows the best accent performance in term of sentence and speaker level correlations. In [6], GOP was used to provide a score at phone level for non-native speech by computing the duration normalized log of posterior probability of the uttered phone given the corresponding acoustic segment. The reliability of the GOP scoring depends heavily on the quality of the native trained-acoustic models used. Furthermore, [7] investigates the linear and non-linear (neural-network) combinations of the confidence scores: log-posterior probability and segment duration scores. The results show improvement at the sentence level with a high degree of correlation when log-posterior and segment duration scores are combined in a non-linear manner

Manuscript received April 17, 2012; revised September 22, 2012, February 08, 2013; accepted March 20, 2013. Date of publication April 12, 2013; date of current version July 11, 2013. This work was supported by the Air Force Research Laboratory (AFRL) under Contract FA8750-12-1-0188 (approved for public release, distribution unlimited), and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

The authors are with the Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX USA (e-mail: abhijeet.sangwan@utdallas.edu; john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2258011

using a neural network. The work in [8] presents automatic accent analysis based on Phonological Features (PFs). Markov models of PFs extracted from native and non-native speech are formulated and employed to develop a statistical measure of accentedness which rates pronunciation of a word on a scale of native like $(+1)$ to non-native like (-1). From the algorithms implemented in the past, it is observed that generally accent scores are computed by building a pronunciation template against which non-native speech can be compared and scored.

Generated accent scores described in previous paragraph show high correlation when compared to human evaluated accent scores. Particularly, in this study, we are interested in the phone level assessment. While a system which generates scores at phone level with high correlation when compared to human evaluated accent scores has been achieved [11], it has been suggested that phone level accent scores alone are not sufficient to give feedback to L2 learners. In fact, they need to be supplemented with information of the pronunciation mismatches [12]. Here, traditional assessment algorithms such as GOP [6] focus on measuring the impact of phone level substitutions, and ignore phone level deletions and insertions. Therefore, it is desirable to build automatic accent assessment systems that are able to score pronunciation mismatches at phone level (i.e., phone level substitutions, deletions, and insertions). In this study, we capture and detect pronunciation mismatches in L2 speech by alignment. Alignment of canonical phone sequence (corresponding to the canonical pronunciation) with L2 speakers spoken phone sequence reveals the exact nature of mismatch in terms of phonetic substitutions, insertions, and deletions. The reliability of accent scores assigned to pronunciation mismatches is crucial towards providing feedback to L2 learners. Here, it is reasonable to suggest that the perceptual impact of different pronunciation mismatches is not equal. In other words, a different number and type of pronunciation mismatches would lead to varying perception of the degree of foreign accentedness (i.e., mild to heavy foreign accent). Hence, automatic accent scoring systems that can incorporate human perception into the scoring paradigm would enhance the value of the feedback provided to L2 learners. Traditional accent scoring algorithms such as the GOP tend to rely on acoustic models for score generation and do not account for perception. In this study, we propose a novel accent scoring technique that can automatically model the perceptual impact of different pronunciation mismatches by extracting patterns of scoring from listener evaluations. In summary, the proposed accent assessment technique in this study first determines pronunciation mismatches in spoken utterances and subsequently uses a perceptual model to score the mismatches to generate an accent score. The algorithm operates in 2 steps: alignment and scoring. In the alignment step, Weighted Finite State Transducers (WFSTs) are employed to capture phone level substitutions, deletions, and insertions by aligning the decoded (spoken) and canonical phone sequences. WFSTs are very versatile and have been utilized in various speech and language processing applications such as speech-to-speech translation, pronunciation modeling, compilation of morphological and phonological rules, and very large-scale dictionary representation [13]–[16]. In the scoring step, the proposed system incorporates the new idea of incorporating L1 perceptual information through the use of Maximum Entropy Model (MEM) which automatically learns the penalty associated with different type of pronunciation mismatches from human evaluation of native and non-native speech. MEM has been successfully applied in the field of Natural Language Processing (NLP) for applications such as part-of-speech (POS) tagging [17], machine translation (MT) [18], [19] and acoustic modeling [20]. To the best of our knowledge, the proposed system in this study is the first automatic pronunciation assessment system which incorporates perceptual information of L1 speaker to assign scores to pronunciation mismatches. While a holistic accent evaluation system should include assessment of tone/pitch, stress, rhythm etc. [21], we only focus on phonetic mismatches in this study. The proposed system is evaluated on isolated words of AE spoken by Native Mandarin Chinese (N-MC) and Native AE (N-AE) from CU-Accent corpus [22]. By conducting exhaustive listener evaluation study, the accent ground truth for the speech samples in the CU-Accent corpus is established. Finally, speech by N-MC and N-AE is scored for accentedness by the proposed system and correlated with the human generated scores. The experimental results demonstrate the feasibility of the proposed approach to provide consistent results when compared with human evaluation. The rest of the paper is organized as follows: In Section II, we discuss the proposed automatic accent assessment systems in detail, namely: WFST and Perceptual-WFST (P-WFST). In Section III, we describe the CU-Accent corpus, listener evaluation, and models training for the proposed accent assessment systems. In Section IV, we present the results and discuss the experiments conducted in this study.

II. Proposed Accent Assessment System

The proposed perceptual WFST (P-WFST) accent assessment technique is shown in Fig. 1. In the system front-end, the acoustic signal is pre-emphasized with a factor of 0.97 and followed by frame analysis using a 25 ms window with a 15 ms shift. Next, 13 dimensional Mel Frequency Cepstral Coefficients (MFCC) features are extracted using a set of 40 triangular filters to simulate the Mel-Scale, along with the delta and delta-delta MFCCs which contribute to a total of 39 dimensions. After MFCC extraction, the acoustic signal is decoded using monophone HMMs. Here, the decoding graph is generated dynamically from the canonical phone sequence with the intention of capturing variability in pronunciation. As shown in Fig. 1, this is accomplished by constructing the decoding graph in a manner that represents the most likely phone-level substitutions, deletions, and insertions as alternate hypotheses to the decoder. Certain substitutions are highly unlikely and therefore not allowed by the decoding graphs (e.g. ($/s/$ to $/aa/$), ($/d/$ to $/ah/$)). Table I shows the complete list of phone substitutions used in the proposed system. The phone substitutions are largely inspired by shared place and manner of articulation. Additionally, some substitutions are hand-crafted based on knowledge of non-native (Mandarin

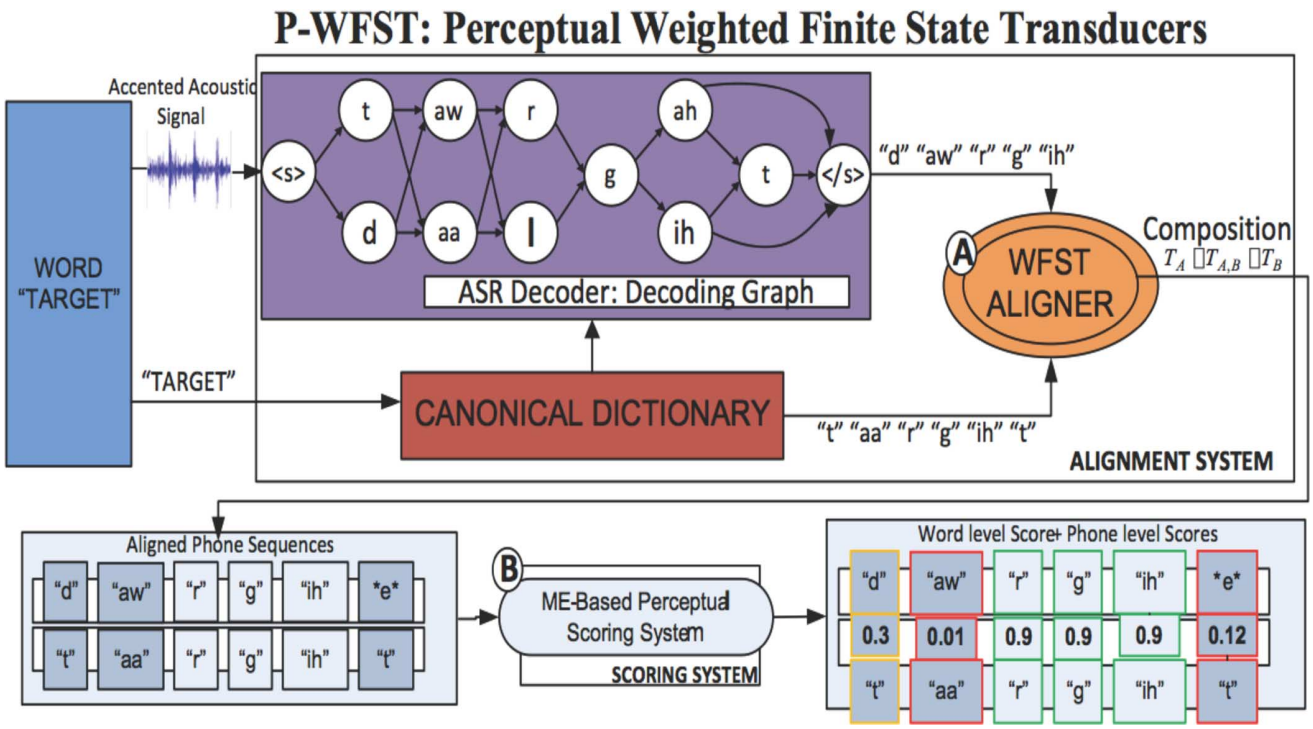

Fig. 1. Proposed automatic accent assessment method uses (Weighted Finite State Transducer) WFST based technique (A) to automatically detect pronunciation mismatches, WFST Scoring System to provide accent scores (B), and (Maximum Entropy) ME based perceptual (C) scoring technique to assign penalties to the pronunciation mismatches, as well as provide perception based accent scores.

TABLE I
PHONEME MAPPING STRATEGY

Allowable Phoneme Mappings					
Stop Consonants		Vowels		Semi Vowels	
Phone	Allowable Mapping	Phoneme	Allowable Mapping	Phoneme	Allowable Mapping
/d/	/d/, /t/, /dh/, /th/	/iy/	/iy/, /ih/	/l/	/l/, /eh/
/t/	/d/, /t/, /k/, /g/	/uw/	/uw/, /uh/	/w/	/w/, /v/
/b/	/b/, /p/, /m/	/eh/	/eh/, /ah/, /ae/	/r/	/r/, /er/
/p/	/p/, /b/, /m/	/ah/	/ah/, /aa/	/y/	/y/, /ih/, /iy/
/g/	/d/, /k/, /g/, /t/	/aa/	/aa/, /ah/, /aw/	Fricatives	
/k/	/d/, /t/, /k/, /g/	/er/	/er/, /r/, /ae/, /ah/	/s/	/s/, /sh/, /z/, /zh/
Dipthongs		/ih/	/iy/, /ih/, /ah/	/z/	/s/, /sh/, /z/, /zh/, /ch/
/ey/	/ey/, /ae/, /ay/, /oy/, /eh/, /ah/	/uh/	/uh/, /uw/, /ah/	/zh/	/zh/, /z/, /s/, /sh/, /jh/
/ow/	/ow/, /ao/	/ae/	/ae/, /ah/, /eh/	/f/	/f/, /v/
/oy/	/oy/, /ow/, /ah/	/aw/	/aw/, /ao/, /aa/, /ah/	/v/	/v/, /f/, /w/
/ao/	/ao/, /ow/, /aa/, /ah/	Nasals		/th/	/th/, /dh/
/ay/	/ay/, /ae/, /ey/, /oy/, /ah/	/m/	/m/, /n/, /ng/	/dh/	/dh/, /th/, /t/, /d/
Affricatives		/n/	/m/, /n/, /ng/	/hh/	/hh/
/jh/	/jh/, /ch/, /zh/, /z/, /dh/	/ng/	/m/, /n/, /ng/	/sh/	/s/, /sh/, /z/, /zh/
/ch/	/ch/, /jh/, /zh/, /sh/, /z/, /s/				

Chinese Speakers in this study) articulation, and typical speech recognition errors.

In the proposed technique, the Viterbi algorithm is employed for decoding and choosing the most likely pronunciation path. As shown in Fig. 1, the decoded and canonical phones sequences are then aligned using a WFST. The alignment reveals the potential phone-level mismatches in the pronunciation in terms of phone substitutions, deletions and insertions. The phonetic mismatches in the pronunciation are then processed by a ME-based perceptual scorer that assigns penalty to each mismatch. The ME-based scorer is trained to assign higher penalty to phonetic mismatches that lead to a higher perception of accent. For example, Fig. 1 shows the areas of pronunciation with the higher penalties assigned by ME scorer. In this study, the output of the WFST scoring system as well as the ME scoring system are used to generate accent scores, and these are termed as the WFST and P-WFST (perceptual WFST) accent scores, respectively.

A. WFST Alignment System

A WFST is a directed graph with weighted arcs, and an input and output label designated on each arch. Each vertex is called a state and two states are assigned as the initial and final states. Transduction in the WFST model represents all possible alignments between decoded and canonical phones sequences. In this study, two separate WFST alignment models are constructed for native and non-native speakers. The input and output to the WFST alignment model are the decoded phone

(q_d) and the canonical phone (q_c) sequences, respectively. The WFST weights $(q_d,\ q_c)$ can be interpreted as the conditional probability of the canonical phone given the decoded phone, $P(q_c|q_d)$. These weights are manipulated through the use of real semiring $(R, +, 0, 1)$

B. EM Weight Training for WFST

The Forward-Backward Expected Maximization (FB-EM) algorithm [23] is used to train the WFST weights. An initial WFST framework is constructed in such a way so that it covers all possible phones mappings. Let $T_{A,B}$ be the WFST alignment model which is trained using the FB-EM algorithm whose initial and final states are the same, and $(A_i,\ B_i)$ is the pair of decoded and canonical phone sequences respectively. For a given sequence pair $(A_i,\ B_i)$, multiple paths through the $T_{A,B}$ are possible. The weights for $T_{A,B}$ weights are initialized such that all phones that follow phone mappings in Table I have a value of 1, otherwise the weights are floored to a significantly small value close to 0.

The FB-EM algorithm consists of two stages: Expectation step and Maximization step. In the expectation step, the weight for each phone mapping is computed for across all sequence pairs $(A_i,\ B_i)$ in the training corpus as follows:

1) Compute all possible alignments of (A_i, B_i) by performing compositions [16],

$$M_i = A_i \circ T_{A,B} \circ B_i, M_i \geq 1. \tag{1}$$

2) Normalize the weights of all paths/alignments so that they sum up to 1, where the probability of a path is defined as,

$$\mathbf{P}(M_i) = \prod_{k=1}^{K} \mathbf{P}(q_{ck}|q_{dk}), \tag{2}$$

where q_{ck} and q_{dk} are the kth canonical and decoded phone in M_i, respectively, $i \geq 1$. The new updated $\mathbf{P}_{\text{new}}(M_i)$ can be calculated as,

$$\mathbf{P}_{\text{new}}(M_i) = \frac{\mathbf{P}(M_i)}{\Sigma_i \mathbf{P}(M_i)}, i \geq 1. \tag{3}$$

3) For each $(q_d,\ q_c)$, count instances of all phone mappings as observed in all alignments M_i of all pairs of sequences $(A_i,\ B_i)$. Each M_i contributes its weight to conditional probability of $(q_d,\ q_c)$ as

$$\mathbf{P}(q_{ck}|q_{dk}) = \Sigma_i N_{M_i} \mathbf{P}_{\text{new}}(M_i), \tag{4}$$

where $i \geq 1$, and N_{M_i} is the number of occurrences of a particular $(q_{dk},\ q_{ck})$ in M_i across all pairs of sequences $(A_i,\ B_i)$. Finally, the probability $\mathbf{P}(q_{ck}|q_{dk})$ is subsequently normalized.

In the Maximization step, the alignment scores are re-computed for all pairs of sequences $(A_i,\ B_i)$ from the product of the updated weights $\mathbf{P}(q_{ck}|q_{dk})$ corresponding to each alignment and normalized such that the total probability of all paths sum to 1. The training iteratively uses (2), (3), and (4) until the weights converge. At termination, the WFST weights capture the frequency of pronunciation mismatches at phone level.

C. Alignment of Decoded-Canonical Phone Sequences

Consider a cascade of FSTs $M_i = T_A \circ T_{A,B} \circ T_B$, where T_A and T_B are the FST of the decoded phones sequence and canonical phones sequence respectively, whose edges have the same input-output labels, then M_i represents all possible alignments between the decoded and canonical phone sequence. The most likely alignment can be estimated as,

$$M^* = \arg\max_{M_i} \mathbf{P}(M_i), \tag{5}$$

and by combining (2) and (5), we obtain,

$$M^* = \arg\max_{M_i} \prod_{k} \mathbf{P}(q_{ck}|q_{dk}). \tag{6}$$

The alignment between decoded and canonical phones sequences consists of the sequence of input-output labels of the WFST resulting from (6). This alignment captures the pronunciation mismatches at the phone level by exposing substitution $(q_{dk},\ q_{ck})$, deletion $(q_{dk} = *e*)$, and insertion $(q_{ck} = *e*)$ events, where $*e*$ represents empty phoneme. For example, the optimal alignment of word "target" is shown in Fig. 1 as the output from composition, $T_A \circ T_{A,B} \circ T_B$. The output is then processed by the WFST accent scoring which is described in the next section.

D. WFST Accent Score

The WFST accent score utilizes the Normalized Delta Log-Likelihood [8] to assign varying degree of "accentedness" score in the range $[-1, +1]$, where -1 represents extreme foreign accent (N-MC) and $+1$ represents native like pronunciation, N-AE. The accent score for a particular word can be computed as:

$$\Lambda = \frac{P(M * T_{\text{AE}}) - P(M * T_{\text{MC}})}{P(M * T_{\text{AE}}) + P(M * T_{\text{MC}})}, \tag{7}$$

where $\Lambda \in [-1, +1]$. $P(M * T_{\text{AE}})$ and $P(M * T_{\text{MC}})$ denote the probabilities of optimal alignment between canonical and the decoded phone sequences when composed with the WFST alignment model (trained on native AE data) and (trained on non-native MC data). While the WFST accent score detects and captures pronunciation mismatches, it does not account for the impact of different pronunciation mismatches on human perception. To overcome this limitation, we develop the ME-based perceptual accent scoring in the next section.

E. ME-Based Perceptual Scoring System

In this study, N-AE perceptual information is incorporated in the proposed pronunciation assessment system through the use of a Maximum Entropy Model (MEM). In the proposed system, MEM learns the manner in which N-AE listeners judge spoken utterances by native and non-native speakers. The learning is achieved by training MEM on data from listener evaluations, where N-AE listeners assign accent scores to utterances spoken by native and non-native speakers. These accent scores are in the continuous range from 0–100. The details of the listener evaluation study are presented in Section 3.2. The goal of MEM is to

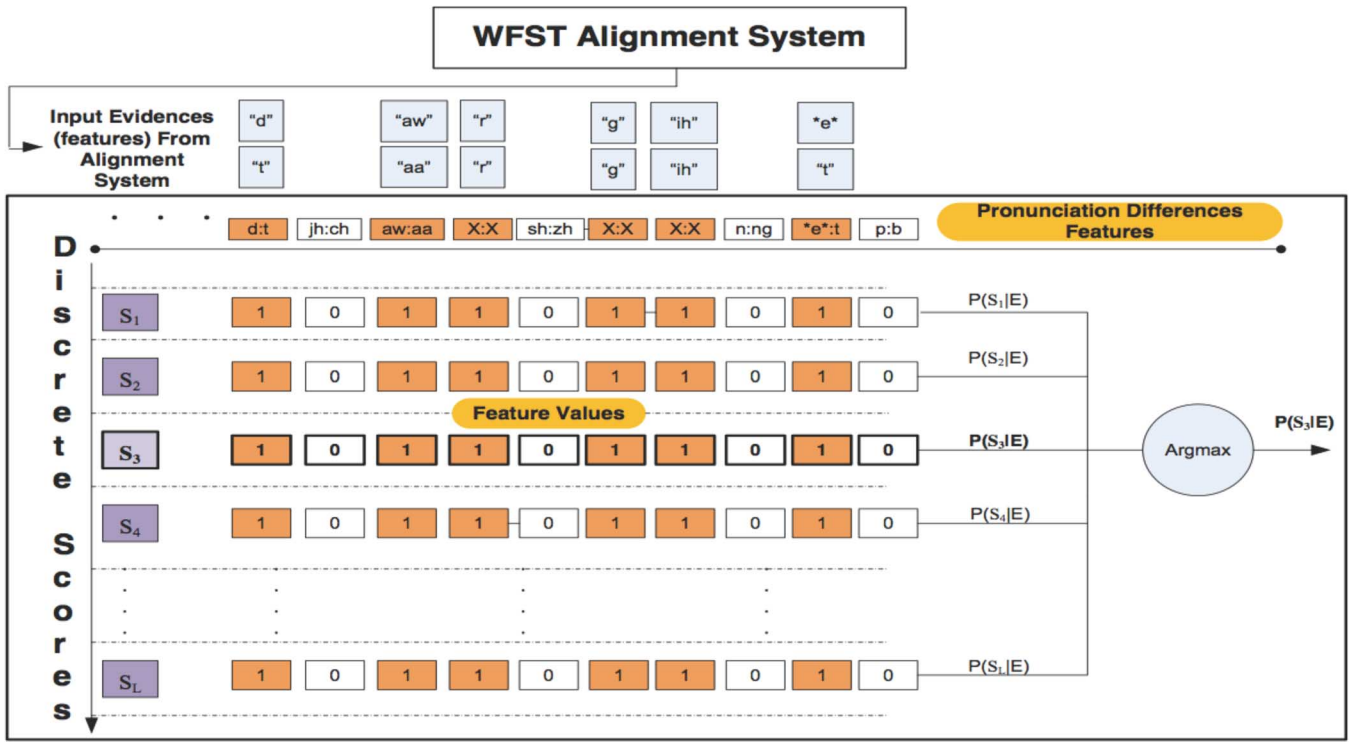

Fig. 2. Proposed Maximum Entropy (ME) scoring system. Pronunciation mismatch features are obtained from WFST alignment. Feature pruning strategy is employed to group non-error pronunciation features, e.g., (r:r), (g:g), (ih:ih), are mapped to (X:X). Feature values are 1 whenever the pronunciation features are observed in the input evidences from WFST alignment system, and 0 when not observed.

model the accent scores (S) for pronunciation assessment conditioned on various evidence observed in the input (pronunciation mismatches), (i.e., substitution (q_{dk}, q_{ck}), deletion ($q_{dk} = *e*$), and insertion ($q_{ck} = *e*$)). Upon training, MEM is able to learn weight parameters which capture the impact of pronunciation mismatches on human perception of accent. In our MEM framework, these evidences are implemented as ME features. Since the MEM framework employed in our study makes use of discrete classification instead of estimation on a continuous scale, the N-AE scores are quantized.

As shown in Fig. 2, the set of evidences used in our MEM system are pronunciation mismatches. In this example, the word considered is "target." The input to the ME scoring system are pronunciation mismatches estimated from the WFST alignment system (i.e., (d:t), (aw:aa), and (*e*:t)). The pronunciation mismatches serve as the input evidences E which are used to compute the conditional probability of accent score S given E:

$$p(S|E) = \frac{1}{Z}\exp(\Sigma_{i=1}^{L}\lambda_i f_i) \tag{8}$$

where L is the total number of all possible pronunciation mismatches at the phone level and Z is a normalization factor to ensure the probability of $p(S|E)$ is bounded by 1. Each ME feature is a binary operator on the evidence, (e.g., if the feature is observed, it produces a value of 1, otherwise it is 0),

$$f_i = \begin{cases} 1 & \text{if the } i\text{th features occurs in } E, \\ 0 & \text{otherwise.} \end{cases}$$

For example, in Fig. 2, only evidences $(d : t)$, $(aw : aa)$, $(r : r)$, $(g : g)$, $(ih : ih)$, and $(*e* : t)$, receive feature value of unity (indicated by coral color box), while the other features are zero (indicated by white color boxes). In Fig. 3, L is the number of discrete score categories used, (e.g., $S_1, S_2, \ldots, S_L$), and the corresponding posterior category probabilities $p(S_1|E), p(S_2|E), \ldots, p(S_L|E)$ can be computed using (8). The most likely accent score S given a set of input evidences E is selected at the end of the process, and this represents the word level accent score.

Furthermore, it is hypothesized that matched phone pronunciations have limited impact on accent perception. Therefore, ME features such as $(aa : aa)$, $(t : t)$, $(d : d)$, $(f : f)$, etc. are reduced by replacing them with a generic feature $(X : X)$. This feature reduction strategy allows the MEM to focus on learning the impact of pronunciation mismatch, and also improves the training quality by reducing the number of model parameters. Additionally, the accent scores acquired from listener evaluation in this study are continuous, and needs to be quantized to discrete scores categories in order to train the MEM. Fig. 3 shows the implementation of continuous-to-discrete score conversion with 4 discrete score categories.

III. Experiments

A. Speech Corpus: CU-Accent

The CU-Accent corpus consists of 179 speakers (72 male and 107 female subjects). The CU-Accent corpus consists of speech utterances spoken by native speakers of American English (AE), Mandarin Chinese (MC), Turkish, Spanish, Thai, Japanese, German, Hindi, and French [22]. Participants in the corpus spoke in English as well as their native language during corpus collection, each speaker was asked to speak (i) 23 isolated words, (ii) 4 sentences in English as well as their native language, and (iii) 1 minute of spontaneous monologue on a

TABLE II
COMPOSITION OF THE CU-ACCENT CORPUS IN TERMS OF ISOLATED WORDS, PHRASES, AND SPONTANEOUS SENTENCES

Speech	Tokens	Repetitions per Token	Words and Phrases
Isolated Words	23	5	aluminum, bird, boy bringing, catch, communication, feet, hear, line, look, pump, root, south, student, target, teeth, there, thirty, three voice, white,
Phrases	8	4	This is my mother. He took my book. How old are you? Where are you going?
Spontaneous	1	1	1 minute monologue on any topic of choice.

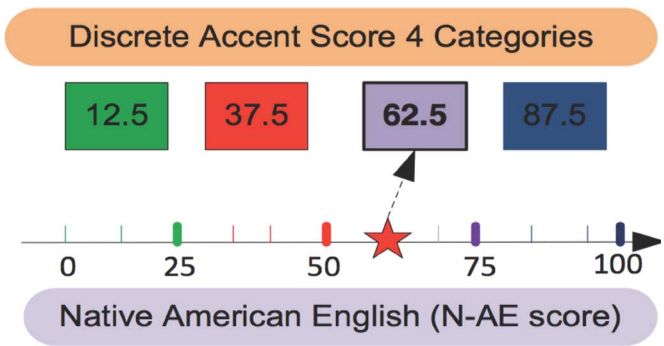

Fig. 3. Illustration of Quantizing Continuous Native American English (N-AE) Score to Discrete Accent Score with 4 categories. Red star on the figure represents the continuous N-AE score which falls within certain range $(50 < N-AE \leq 75)$, and being quantized to one of the discrete score categories (62.5).

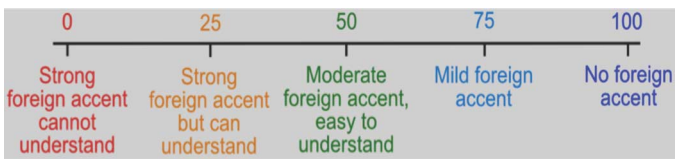

Fig. 4. Continuous Scale from 0-to-100 used by N-AE listeners to rate accents during the Listener Evaluation.

subject of their choosing. This collection protocol was repeated to include session variability in the corpus. The isolated words and phrases used in data collection are shown in Table II. The data for CU-Accent data was acquired using a telephone based dialog system and the speech samples were digitized at 16 bits per sample PCM (Pulse Coded Modulation) with a sample rate of 8 kHz. The words and phrases in the CU-Accent corpus are known to be accent sensitive for non-native speakers of AE in term of phonetic structure and transitions [24]. This corpus, as well as the words/sentence structure, has been used for analysis of automatic and accent classification in the past [8]–[10]. In this study, we focus on isolated words spoken by N-AE and N-MC speakers.

B. Listener Evaluation

In this study, 50 N-AE listeners were asked to rate isolated words in English spoken by N-AE and N-MC speakers from the CU-Accent Corpus. The listener evaluation was divided into 3 sessions. Table III presents the composition of listeners and speakers involved in the sessions along with the materials used in each session. Each speech token used in the evaluation was built by concatenating a particular isolated word three times. For example, 3 repetitions of the word thirty were combined with pauses in-between to form a token (,i.e., thirty - pause- thirty

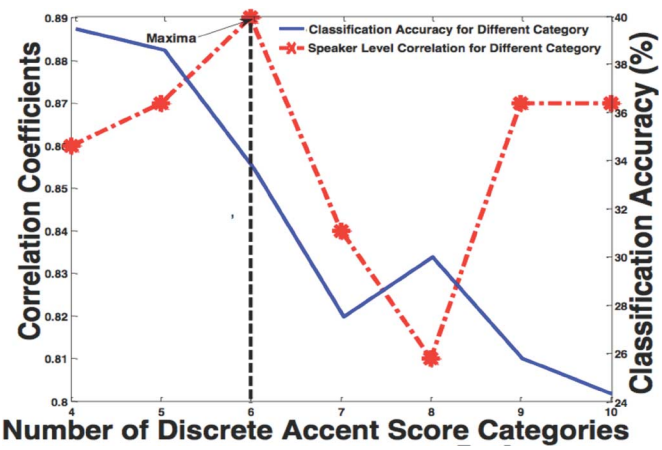

Fig. 5. Classification accuracy and speaker level correlation for (Perceptual Weighted Finite State Transducer) P-WFST proposed system for 7 discrete score categories (4-to-10), evaluated on Set A (combination of N-AE and N-MC). P-WFST with 6 discrete score category obtained the most desirable parameters (correlation: 0.89 significant at $p < 0.001$, accuracy: 33.66%).

-pause- thirty). This concatenation was performed in order to increase the speech material presented to the human listeners. The listeners gave one score for one token [8]. The N-AE listeners were asked to rate the accent level in the speech token on a continuous scale from 0-to-100, as shown in Fig. 4. Here, the score 0 represents a heavy foreign accent (N-MC), and 100 represents no-perceived-accent (N-AE), respectively. The procedure of the listener evaluation used in this study is similar to the one in [8] and [25]. In our experiments, 5 N-AE listeners scores from sessions 1 and 2 are used as ground truth against which the performances of the automatic accent assessment algorithms are validated and this set is referred to as the Testing Set. The rest of the listener scores from the evaluation are used as the training data for building Maximum Entropy (ME) models.

C. Evaluation Analysis

A total of 10 N-AE listeners scores from both Sessions 1 and 2 are used to measure the performance of automatic accent assessment algorithms using correlation of machine and listener scores. In this study, we also measure the inter-rater correlation at word and speaker level to measure the consistency of the listeners scores. The inter-rater correlation serves two purposes: (i) it establishes the capability of native listeners to effectively assess accents of non-native speakers using isolated utterances of a small set of words, and (ii) it also demonstrates the extent of consensus among native listeners with respect to assessing foreign accents. The inter-rater correlation at the word

TABLE III
COMPOSITION OF LISTENER EVALUATION CONDUCTED IN THIS STUDY WITH TOTAL OF 50 LISTENERS FOR 3 SESSIONS. THERE IS OVERLAP BETWEEN SPEAKERS IN ALL SESSION 1, 2 AND 3, BUT NO OVERLAP AMONG LISTENERS

Session	Number of AE-Listeners	Number of Speakers		Number of tokens/session	Isolated Words Involved
		N-AE	N-MC		
1	21	2	7	207	aluminum, bird, boy, bringing, catch, change, communication, feet, hear, line, look, pump, root, south, student, target, teeth, thirty, three, voice white, would,
2	19	3	6	207	aluminum, bird, boy, bringing, catch, change, communication, feet, hear, line, look, pump, root, south, student, target, teeth, there, thirty, three, voice, white, would
3	10	4	18	103	target, three, thirty, hear

TABLE IV
WORD AND SPEAKER LEVEL INTER-RATER CORRELATIONS FOR 5 N-AE LISTENERS—SESSION 1. THE CORRELATIONS ARE SIGNIFICANT AT $p < 0.001$ LEVEL

Correlation Type	Level	Listener's ID					Average
		1	2	3	4	5	
Inter-rater	Word	0.86	0.83	0.84	0.79	0.77	0.82
Inter-rater	Speaker	0.97	0.98	0.99	0.95	0.96	0.97

TABLE V
WORD AND SPEAKER LEVEL INTER-RATER CORRELATIONS FOR 5 N-AE LISTENERS—SESSION 2. THE CORRELATIONS ARE SIGNIFICANT AT $p < 0.001$ LEVEL

Correlation Type	Level	Listener's ID					Average
		6	7	8	9	10	
Inter-rater	Word	0.68	0.64	0.65	0.62	0.69	0.66
Inter-rater	Speaker	0.94	0.95	0.84	0.93	0.80	0.87

level is computed by evaluating the correlation coefficient between the scores given by a single N-AE listener against the average scores of the rest of the N-AE listeners using all isolated utterances of that word. Furthermore, the inter-rater correlation at the speaker level is obtained by computing the correlation between average words scores given by a single N-AE listener and the average of scores of the rest of the N-AE listeners using all utterances of the speaker. Tables IV and V show the inter-rater correlation at word and speaker levels for Sessions 1 and 2. The inter-rater correlation at word-level is observed to vary between 0.86 (Listener 1) and 0.62 (Listener 9). Additionally, the inter-rater correlation at speaker-level varies from 0.99 (Listener 3) to 0.8 (Listener 10). The higher agreement between native listeners at the speaker level versus word level is expected as listeners are exposed to more utterances at the speaker level and their judgment is averaged over more observations (both factors contributing towards lowering the error in judgment). On the average, the inter-rater correlation is observed to be 0.76 and 0.92 at the word and speaker levels. The high correlations demonstrate the capability of native listeners to assess accents with good accuracy and consistency with a relatively small vocabulary set (i.e., 1–23 words). Additionally, the inter-rater correlation at the speaker level (0.92) also suggests a milestone in terms of performance for automatic assessment techniques.

IV. MODELS TRAINING

A. HMM Based Mono Phone Models

We trained 128 mixtures context independent monophone HMMs to build our constraint monophone decoder which covers 39 AE phones: closures $(/p/, /b/, /d/, /g/, /t/, /k/)$, affricatives $(/jh/, /ch/)$, fricatives $(/s/, /sh/, /z/, /zh/, /f/, /th/, /v/, /dh/)$, nasals $(/m/, /n/, /ng/)$, semivowels $(/r/, /w/, /y/, /hh/, /l/)$, and vowels $(/iy/, /eh/, /ey/, /ae/, /aa/, /aw/, /ay/, /ao/, /oy/, /ow/, /uw/, /uh/, /er/, /ah/, /ih/)$, trained on 10,060 AE utterances (5.5 hours of training data) from the CU-Accent corpus, which includes both spontaneous sentences and isolated words. The monophone HMMs were trained using SPHINX. The monophone HMMs are then used to decode the incoming speech into phone sequences.

Next, to ensure that WFST alignment models capture the variability in pronunciation for each group (N-AE and N-MC), sufficient amount of data is needed for training. Therefore, we use 12,364 (55 from N-AE speakers) and 13,654 (24 from N-MC speakers) isolated words from the CU-Accent corpus for AE and MC models respectively. For each group, pairs of decoded and canonical phone sequences are generated for all isolated words. The decoded phone sequences are obtained by passing the isolated words to the constraint monophone decoder obtained in the previous paragraph, while the canonical phone sequences are obtained from a dictionary. Using this training data and the Carmel Toolkit [26], the FB-EM training is performed to generate WFST models for N-AE and N-MC. The Carmel Toolkit is also used to perform WFST composition for aligning decoded and canonical phone sequences.

For the proposed MEM scoring system, the training data is acquired from 40 N-AE listeners from Sessions 1, 2 and 3 of listener evaluation. For all word tokens used in these 3 sessions, the pronunciation mismatches are obtained from the WFST aligner, and are used as MEM features. Feature reduction strategy is then applied to reduce the number of ME features used from 98 to 64. Next, the accent scores acquired from listener evaluation are discretized. For example, the continuous accent score is assigned to 4 discrete score categories 0–25, 25–50, 50–75, and 75–100 and reassigned average category score, (i.e., 12.5, 37.5, 62.5, and 87.5).

Using the procedure described, we implemented seven different MEMs by using 4 and up to 10 discrete categories. Using a larger number of discrete categories allows more resolution in the accent scores. However, more categories also increases classification error for adjacent categories in the MEM system. In this study, our goal is to use the optimum number of categories such that the correlation between automatic and human assigned accent scores is maximized. Using this training data, the Maxent Toolkit [27] is then utilized to train the proposed MEM system to predict the correct discretized accent score from the pronunciation mismatch features.

B. Baseline System GOP

The Goodness Of Pronunciation (GOP) algorithm can compute accent scores at sentence, word, and phone levels [4], [7]. In this study, the implementation of GOP algorithm is similar to the one in [6]. We briefly review the algorithm. GOP algorithm calculates and estimates the posterior probability of a phone:

$$\mathrm{GOP} = \frac{1}{N}(\log(\mathbf{P}(O|p)) - \max_{i \in M} \log(\mathbf{P}(O|p_i))) \qquad (9)$$

where p is the target phone, p_i is the ith phone, O is the acoustic observation, M is the set of all possible phones, and N represents the duration of specific target phone in number of frames. In this study, the monophone HMMs (training described in Section IV.A) are used for GOP score estimation.

V. Results and Discussions

We have used the CU-Accent Corpus for evaluating the proposed accent assessment algorithms, namely WFST (Weighted Finite State Transducers) and P-WFST (Perceptual Weighted Finite State Transducers) against the GOP (Goodness Of Pronunciation). In this study, all experimental analysis is performed on the Test Set described in Section III.B. In order to assess the effectiveness of WFST, P-WFST, and GOP algorithms in measuring accent, two data sets are created and their correlations for the data sets computed separately. One set (Set A) consists of N-AE and N-MC speakers while the other set (Set B) consists of N-MC speakers only. Both sets A and B are derived from the testing set (described in Section III.B). Set A is equivalent to the testing set, and contains a total of 2070 tokens. Since set B only contains N-MC speakers, it is a subset of set A and contains 1495 tokens.

Table VI shows the speaker and word level correlations for Set A and Set B, for all 7 discrete accent score categories of the proposed P-WFST system. It is observed from the table that the word level correlation show less fluctuation across various P-WFSTs score categories for both Sets A (~ 0.3) and B(~ 0.28), However at the speaker level, the correlation coefficient obtained from evaluating Set B has larger fluctuation (as large as 0.18 difference between the largest and the smallest values). It is also observed that, P-WFST with 6 discrete accent score categories delivers the highest speaker level correlation scores for native as well as non-native speakers.

It is important to note that the systems that deliver the highest MEM classification accuracy is not the same as the system that delivers the highest correlation between human and P-WFST generated accent scores. This fact is examined below. In Fig. 5, the accuracy of MEM classification and speaker level correlation for all 7 categories are shown for Set A. It is observed that as the number of categories increase, the classification accuracy of the model decreases (shown in solid blue line). However, the speaker level correlation initially increases and then decreases (shown in dashed red line). Here, P-WFST with 6 categories delivers the highest speaker level correlation (0.89). Therefore, P-WFST with 6 discrete accent score categories is used as the default P-WFST system for the remainder of this study.

TABLE VI
Word and Speaker Level Correlation Coefficient for P-WFST System Evaluated on Set A and Set B, for All Discrete Accent Score Categories. The Correlations Are Significant at $p < 0.001$ Level

Algorithm		Correlation Coefficient			
		Word		Speaker	
	Category	Set A	Set B	Set A	Set B
P-WFST	4	0.33	0.28	0.86	0.68
	5	0.34	0.25	0.87	0.71
	6	0.34	0.31	**0.89**	**0.86**
	7	0.33	0.26	0.84	0.77
	8	0.33	0.28	0.81	0.68
	9	0.34	0.26	0.87	0.81
	10	0.35	0.27	0.87	0.71

TABLE VII
Correlation Between Human and Machine as Well as Human and Human (Inter-Rater) Accent Scores, Set A Consists of N-AE and N-MC Data and Set B Consists of N-MC Data Only. The Correlations Are Significant at $p < 0.001$ Level

Algorithm	Correlation Coefficient			
	Word		Speaker	
	Set A	Set B	Set A	Set B
Human	0.73	0.60	0.95	0.81
WFST	0.31	0.15	0.88	0.63
P-WFST	0.34	**0.31**	0.89	**0.86**
Baseline-GOP	0.47	0.27	0.89	0.75

The next analysis compares the proposed accent assessment systems (WFST and P-WFST) against baseline GOP and human inter-rater evaluation in terms of word and speaker level correlations. Table VII shows the results of this comparison for Sets A and B. It is observed that the WFST achieves a correlation of 0.88 at the speaker level, and 0.31 at the word level, while the P-WFST system reaches a higher speaker level correlation of 0.89, while attaining a word level correlation of 0.34. On Set A, P-WFST system matches GOP performance at speaker level (0.89) and is lower than GOP performance at the word level (0.47). On the other hand, for Set B (i.e., only N-MC speaking AE), WFST correlations at the word and speaker levels are 0.15 and 0.63, respectively, and P-WFST system attains a higher correlation of 0.31 (word level) and 0.86 (speaker level). Here, P-WFST outperforms GOP performance at both the word and speaker level correlation (i.e., 0.27 and 0.75, respectively).

The results obtained in Table VII show that speaker level correlations are consistently higher than the word level correlations. This result signifies that the average score based on several words is a more reliable measure of accent than the scores

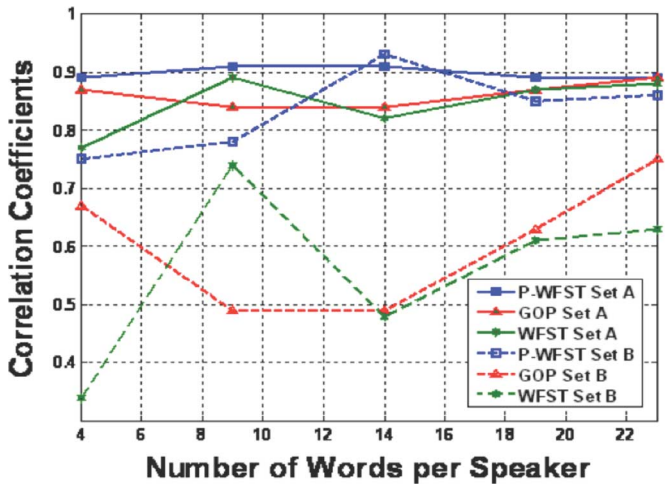

Fig. 6. Speaker level correlation number of words WFST (Weighted Finite State Transducer), P-WFST (Perceptual-WFST) and GOP (Goodness of Pronunciation) for Set A and Set B. Set A consists of N-AE and N-MC data and Set B consists of N-MC data only. The correlations are significant at $p < 0.001$ level.

based on a single word. This is more so for automatic assessment algorithms than humans. Furthermore, both human and algorithm correlation performances drop on Set B. Since accent classification (separating N-MC from N-AE speakers) is built into the accent assessment in Set A, we believe the task is inherently easier than accent assessment of Set B which contains only MC non-native speakers speaking AE. In other words, it is easier for humans and automatic algorithms to rate native speakers (consistent high proficiency) as opposed to rating non-native speakers (more likely to be distributed across a range of proficiency levels).

The improved performance of P-WFST on Set B is particularly notable since this set consists of non-native speakers only. The increased agreement between P-WFST and human listeners shows that P-WFST can identify different proficiency groups within non-native speakers more effectively. Hence, the P-WFST can be a more reliable and accurate measurement of accent.

We also conducted an experiment to investigate the relationship between number of words used to compute accent scores and the speaker level correlation performance of WFST, P-WFST, and GOP. On Set A, we observe from Fig. 6 that by averaging the accent scores of 4 words only, P-WFST reaches a higher correlation of 0.89 compared to WFST and GOP (0.77 and 0.87, respectively). Additionally, as the number of words increases, the overall algorithm performance for WFST, P-WFST and GOP also increase. On Set B, the P-WFST system reaches a higher correlation of 0.75 compared to that of WFST and GOPs (0.34 and 0.69, respectively) by using accent scores from 4 words. As seen in Set A, P-WFSTs correlation performance increases with an increase in the number of words used. However, the WFST and GOP performances fluctuate as the number of words increase. This observation suggests that WFST and GOP need more samples to reliably estimate accent, while P-WFST is able to estimate accent with fewer samples. In fact, it is observed that the P-WFST achieves high performance with little data (7–8 words are sufficient to provide

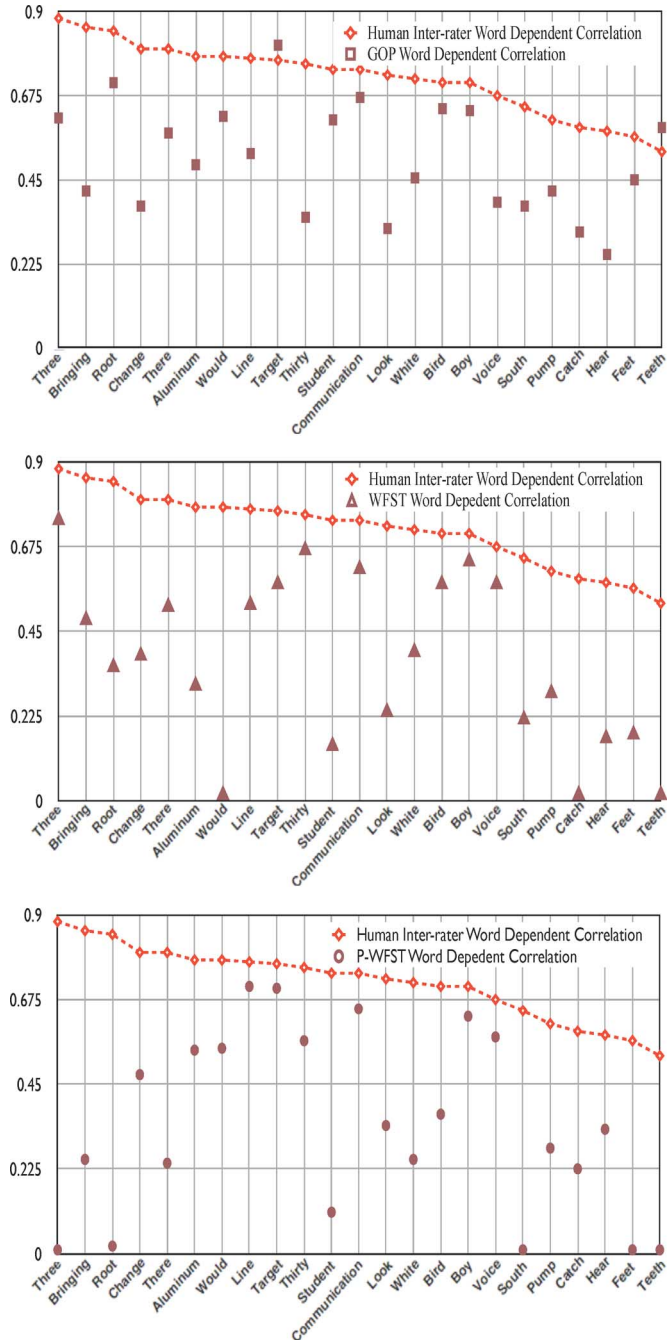

Fig. 7. Word-dependent correlation evaluated on Set A (N-AE and N-MC) with 23 isolated words from CU-Accent corpus. The correlations are significant at $p < 0.001$ level.

accurate measurements). We believe that this stems from the unique approach that P-WFST applies to accent measurement (i.e., penalty assignment to pronunciation mismatches).

An interesting experiment to assess word-dependent correlations for machines and human on Set A is conducted. From Fig. 7, we observe that the words *target*, *communication*, and *boy* exhibit high correlation score agreement for both machine and human, and therefore are suitable for use in accent assessment. On the other hand, the words: *catch*, *feet*, and *hear* possess low correlation which reflects on both human and machines inability to assess accent using these three words. High inter-rater correlation and low human-machine correlation is observed for

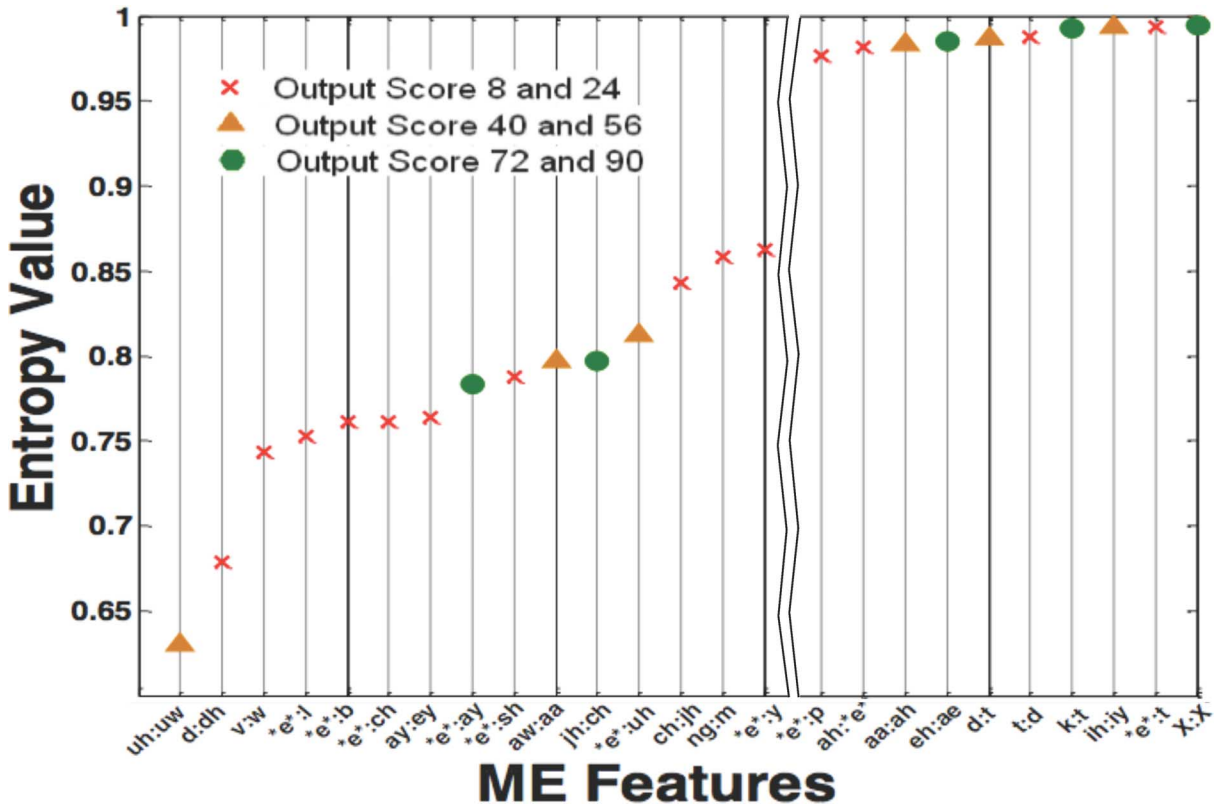

Fig. 8. Maximum Entropy (ME) features and their weights are listed from left to right in the ascending order of entropy values. These 25 features consist of 15 features which inherit lowest entropy and 10 features which inherit highest entropy values.

the words: *bringing*, *change*, and *look*. The algorithms are less effective in assessing accent using these 3 words, while on average, human listeners can judge their accent structure fairly easily. When compared to the algorithms used in this study, humans have access to additional information (e.g., prosody) to judge accent; and the addition of this knowledge would improve automatic algorithm performance as well. Future work will consider developing such a holistic approach where information from multiple sources such as phones, prosody, etc. are combined for an overall accent assessment.

Finally, we analyze the Maximum Entropy (ME) features which have been trained on listener evaluation data described in Section 3.2. Recall, these ME features are pronunciation mismatches obtained from the WFST alignment model. We are interested in finding the most powerful accent assessment features. In the P-WFST, the most powerful features would also be the most discriminative features for MEM classification. Here, the discriminativeness is computed by calculating the entropy value for each feature, (i.e., entropy of the conditional probability distribution of discrete accent class given the input pronunciation mismatch feature). The interpretation here is that the lower entropy features would be more discriminative and higher entropy features would be less discriminative. In Fig. 8, the features are listed from the lowest to highest entropy values. Let us consider the 6 lowest entropy features, namely: (uh:uw), (d:dh), (v:w), (*e*:l), (*e* :b), and (*e*:ch). The feature (uh:uw) inherits a very discriminative characteristic because of the very low feature entropy that it possesses, particularly, discriminative toward mid range scores (40 and 56). The other 5 features, (d:dh), (v:w), (*e*:l), (*e* :b), and (*e*:ch), are more discriminative toward low range scores (8 and 24). Now, let us analyze the features which inherit the highest values of entropy. This high value of entropy signifies the non-discriminative characteristics of these features. It is expected that the non-error pronunciation feature, (e.g., (X:X)), inherits a very high entropy value and more discriminative toward high range (72 and 90) which signifies less significance of the feature impact on human perception of accent.

VI. Conclusions and Future Work

In this study, a new approach (P-WFST) towards accent assessment that relies on two important steps: (i) detecting pronunciation mismatches (substitutions, deletions, and insertions), and (ii) assigning perceptually motivated penalties to the pronunciation mismatches has been proposed. In particular, a Weighted Finite State Transducer (WFST) based technique is used to detect pronunciation mismatches in speech. Additionally, a Maximum Entropy (ME) based technique is employed to automatically learn pronunciation mismatches penalties from human judgment of accent. The proposed system is evaluated on AE spoken by Native American English (N-AE) and Native Mandarin Chinese (N-MC) speakers from the CU-Accent Corpus. The experimental results showed that: (i) the P-WFST based system achieved consistent correlation at speaker and word levels (0.89 and 0.34 respectively) and outperformed GOP by 14.8% when evaluated on non-native speakers only, (ii) With only 4 words, P-WFST based system is able to achieve higher correlation than GOP.

REFERENCES

[1] J. Logan, S. Lively, and D. Pisoni, "Training japanese listeners to identify english /r/ and /l/: A first report," *J. Acoust. Soc. Amer. (JASA)*, vol. 89, pp. 874–886, 1991.

[2] M. Goudbeek, A. Cutler, and R. Smith, "Supervised and unsupervised learning of multidimensionally varying non-native speech category," *Elsevier: Speech Commun.*, vol. 50, pp. 109–125, 2008.

[3] S. Y. Yoon, M. H. Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Proc. Interspeech'10*, , 2010, pp. 614–617.

[4] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. ICSLP'96*, , 1996, pp. 1457–1460.

[5] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Elsevier: Speech Commun.*, vol. 30, no. 2–3, pp. 83–93, 2000.

[6] S. Witt, "Use of speech recognition in computer assisted language learning," Ph.D. dissertation, Univ. of Cambridge, Cambridge, U.K., 1999.

[7] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. ICASSP'97*, 1997, vol. 2, pp. 1471–1474.

[8] A. Sangwan and J. H. L. Hansen, "Automatic analysis of mandarin accented english using phonological features," *Elsevier: Speech Commun.*, vol. 54, no. 1, pp. 40–54, Jan. 2012.

[9] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent," *J. Acoust. Soc. Amer. (JASA)*, vol. 102, no. 1, pp. 28–40, Jul. 1997.

[10] L. M. Arslan and J. H. L. Hansen, "Language accent classification in american english," *Elsevier: Speech Commun.*, vol. 18, no. 4, pp. 92–95, Apr. 1997.

[11] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, 1997.

[12] H. C. Liao, J. C. Chen, S. C. Chang, Y. H. Guan, and C. H. Lee, "Decision tree based tone modeling with corrective feedbacks for automatic mandarin tone assessment," in *Proc. Interspeech'10*, , 2010, pp. 602–605.

[13] F. Casacuberta, D. Llorens, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, and J. M. Vilar, "Speech-to-speech translation based on finite-state transducers," in *Proc. ICASSP'01*, vol. 1, pp. 613–616.

[14] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Elsevier: Speech Commun.*, vol. 46, no. 2, pp. 189–203, 2005.

[15] M. Mohri, "Finite-state transducers in language and speech processing," *Comput. Linguist.*, vol. 23, no. 2, pp. 269–311, 1997.

[16] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 69–88, 2002.

[17] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proc. Conf. Empir. Meth. Nat. Lang. Process.*, 1996, pp. 133–142.

[18] L. Gu, Y. Gao, L. Fu-Hua, and M. Picheny, "Concept-based speech-to-speech translation using maximum entropy models for statistical natural concept generation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 2, pp. 377–392, Mar. 2006.

[19] A. Berger, S. D. Pietra, and V. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–72, 1996.

[20] H.-K. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 873–881, May 2006.

[21] J. C. Chen, J. S. R. Jang, and T. L. Tsai, "Automatic pronunciation assessment for mandarin chinese: Approaches and system overview," *Comput. Linguist. Chinese Lang. Process.*, vol. 12, no. 4, pp. 443–458, 2007.

[22] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 634–646, Mar. 2006.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. B, pp. 1–38, Jun. 1977.

[24] F. Chreist, *Foreign Accent*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1964.

[25] J. E. Flege, "Factors affecting degree of perceived foreign accent in english sentences," *J. Acoust. Soc. Amer. (JASA)*, vol. 84, no. 1, pp. 70–79, 1988.

[26] J. Graehl, Carmel Tool, [Online]. Available: http://www.isi.edu/natural-language/licenses/carmel-license.html

[27] Z. Le, Maxent-Toolkit, [Online]. Available: http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Freddy William (S'09) was born in Jakarta, Indonesia on August 26, 1986. He was awarded the degree of Bachelor of Science in Electrical Engineering with highest Latin Honor, Summa Cum Laude, in May 2009 at University of Texas at Dallas (UTDallas). He completed a summer internship at Alcatel-Lucent in the area of network element analysis. He continued his research in the area of signal and speech processing at UTDallas, and completed his M.S. Degree in electrical engineering in June 2011.

Abhijeet Sangwan (S'04–M'09) was born in Mysore, India. He received his B.E., M.A.Sc. and Ph.D. degrees in electrical engineering from Visveswaraiah Technological University (VTU), Karnataka, India in 2002, Concordia University, Montreal, Canada in 2006, and The University of Texas at Dallas, Texas, U.S.A., respectively. During 2002–2003, he worked for MindTree Consulting where he designed and developed enterprise datawarehouse systems for Unilever. He interned with the Human Language Technologies Group at IBM's T.J. Watson Research Center, Yorktown Heights in 2008. From 2009, he has been a part of The Center for Robust Speech Systems (CRSS) at The University of Texas at Dallas. His research interests include Automatic Speech Recognition (ASR), Keyword Recognition, Automatic Sentiment Analysis, Automatic Accent Assessment, and Language Identification Systems.

John H.L. Hansen (S'81–M'82–SM'93–F'07) received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, N.J. in 1982. He joined University of Texas at Dallas (UTDallas), Erik Jonsson School of Engineering and Computer Science in the fall of 2005, where he is presently serving as Jonsson School Associate Dean for Research, as well as Professor of Electrical Engineering and also holds the Distinguished University Chair in Telecommunications Engineering. He previously served as Department Head of Electrical Engineering (2005–2012), overseeing a $+4$x increase in research expenditures ($4.5 M to $22.3 M) with a 20% increase in enrollment and the addition of 18 T/TT faculty, growing UTDallas to be the 8th largest EE program from ASEE rankings in terms of degrees awarded. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). At UTDallas, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Dept. Chairman and Professor of Dept. of Speech, Language and Hearing Sciences (SLHS), and Professor of the Dept. of Electrical & Computer Engineering, at Univ. of Colorado Boulder (1998–2005), where he co-founded and served as Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. In has been named IEEE Fellow (2007) for contributions in "Robust Speech Recognition in Stress and Noise," named Inter. Speech Communication Association (ISCA) Fellow (2010) for contributions on research for speech processing of signals under adverse conditions, and received The Acoustical Society of America's 25 Year Award (2010)—in recognition of his service, contributions, and membership to the Acoustical Society of America. He is serving as Past TC-Chair (served as Technical Committee Chair: 2010–12) and Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (2005–08; 2010–13; elected IEEE SLTC Chairman for

2011–2012), and elected ISCA Distinguished Lecturer (2012). He has also served as member of the IEEE Signal Processing Society Educational Technical Committee (2005–08; 2008–10). Previously, he served as The Technical Advisor to the U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/06), Associate Editor for IEEE TRANS. SPEECH & AUDIO PROCESSING (1992–99), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–03). He has also served as guest editor of the Oct. 1994 special issue on Robust Speech Recognition for IEEE TRANS. SPEECH & AUDIO PROCESSING . He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–03), and is serving as a member of the ISCA (Inter. Speech Communications Association) Advisory Council. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 59 Ph.D./M.S. thesis candidates (27 Ph.D., 32 M.S./M.A.), was recipient of The 2005 University of Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 459 journal and conference papers and 11 textbooks in the field of speech processing and language technology, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), co-editor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and lead author of the report "The Impact of Speech Under'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ISCA Interspeech-2002, Sept. 16–20, 2002, and Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.