Using Human Perception for Automatic Accent Assessment

Freddy William, Abhijeet Sangwan, and John H. L. Hansen[1]

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
{fxw061000, abhijeet.sangwan, john.hansen}@utdallas.edu

10.21437/Interspeech.2011-642

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. At the alignment step, the speech utterance is processed using a Weighted Finite State Transducer (WFST) based technique to automatically estimate the pronunciation errors. Subsequently, in the scoring step a Maximum Entropy (ME) based technique is employed to assign perceptually motivated scores to pronunciation errors. The combination of the two steps yields an approach that measures accent based on perceptual impact of pronunciation errors, and is termed as the Perceptual WFST (P-WFST). The P-WFST is evaluated on American English (AE) spoken by native and non-native (native speakers of Mandarin-Chinese) speakers from the CU-Accent corpus. The proposed P-WFST algorithm shows higher and more consistent correlation with human evaluated accent scores, when compared to the Goodness Of Pronunciation (GOP) algorithm.

Index Terms: automatic accent assessment, pronunciation scoring, Finite State Transducers, Maximum Entropy

1. Introduction

Automatic pronunciation assessment systems has received a lot of attention where algorithm based on Hidden Markov Model (HMM) log-likelihood scores, segment classification error scores, segment duration scores, syllabic timing scores [1], [2], the linear and non-linear combination of the confidence scores [3], and the GOP (Goodness Of Pronunciation) measure [4] have been proposed.

In this study, we propose a new approach towards modeling accent by splitting the assessment into 2 steps: alignment and scoring. In term of phones sequences, accented pronunciation can differ from native canonical pronunciation, leading to 3 different types of pronunciation errors: substitution, deletion, and insertion of phones. Traditional assessment algorithms such as GOP focus only on measuring the impact of substitution, while ignoring deletion and insertion errors. In order to address this issue, the proposed system employs Weighted Finite State Transducers (WFST) to capture phone substitution, deletion, and insertion by aligning the decoded and canonical phone sequences. Furthermore, different phone substitutions, deletions, and insertions can be expected to have a different impact on perception of accent. Traditional assessment algorithms such as GOP do not employ perception in scoring. In this study, we propose a Maximum Entropy (ME) based technique that can automatically learn the penalty associated with different types of pronunciation errors from human evaluation of native and non-native accents. In this manner, the proposed assessment strategy accounts for substitution, deletion, and insertion using the alignment process and incorporate perception using the scoring method. The combination of the alignment and scoring techniques is termed as the Perceptual-WFST (P-WFST).

The proposed system is evaluated on American English (AE) spoken by Native AE (N-AE) speakers as well as Native Mandarin Chinese (N-MC) speakers from CU-Accent corpus [7]. When evaluating a combination of N-AE and N-MC speaker data, the proposed P-WFST technique matches GOP performance in terms of correlation with human scores. However, P-WFST shows a higher degree of correlation with human scores than GOP when evaluating on N-MC speakers alone (14.8% higher than GOP). Additionally, speaker level correlation with varying number of words is investigated to analyze the performance of machine as a function of number of words used for assessment. Finally, we also conduct word-dependent correlation experiments to analyze which words are best suited for accent analysis.

2. Proposed Accent Assessment System

The P-WFST accent assessment technique is shown in Fig. 1. As shown in the figure, the technique first decodes the acoustic signal using a standard MFCC-based ASR decoder that utilizes monophone HMMs. Here, the decoding graph is generated dynamically from the canonical phone sequence with an intention of capturing the variability in pronunciation (or pronunciation errors). As shown in Fig.1, this is accomplished by constructing the decoding graph in a manner that presents most likely phone-level substitution, deletion, and insertion as alternate hypothesis to the decoder. Certain substitutions are highly unlikely and therefore not allowed by the decoding graphs, e.g. as (/s/→/aa/), (/d/→/ah/). Additionally, the phone mappings can also be handcrafted by using the knowledge of articulator traits of the target non-native speaker group, e.g. N-MC: (/l/→/r/). In this study, a lookup table is employed for this purpose and a few example entries are shown in Table 1.

[1] This project was funded by AFRL through a subcontract to RADC, Inc. under FA8750-09-C-0067 (Approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

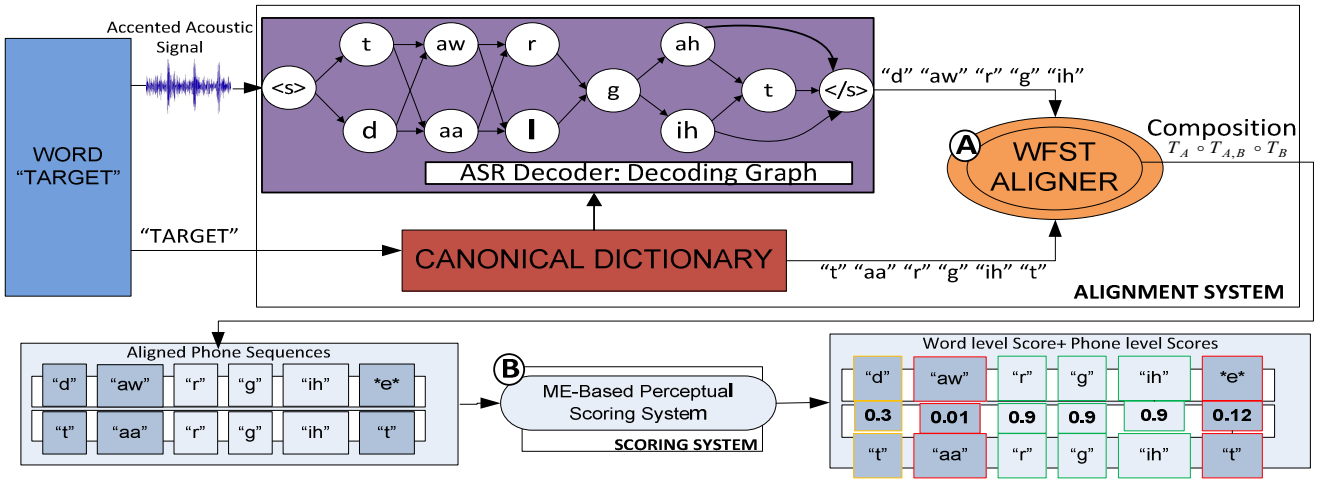

Figure 1: *Proposed automatic accent assessment method uses (Weighted Finite State Transducers) WFST based technique (**A**) to automatically detect pronunciation errors and (Maximum Entropy) ME based perceptual (**B**) scoring technique to assign penalties to the pronunciation errors.*

In the proposed technique, 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 WFST. The pronunciation errors obtained from this alignment are used as input features to ME model (MEM) which utilizes its features' weights to assign penalties to the pronunciation errors. Fig.1 shows the areas of pronunciation with the higher penalties assigned by the proposed system. Finally, MEM also estimates the most likely word-level accent score based on these ME weights.

Table 1. *Example of Phone Mapping Strategy*

Phones	Phone Mappings
/ch/	/ch/ /jh/ /zh/ sh/ /z/ /s/
/n/	/n/ /ng/ /m/
/ey/	/ey/ /ae/ /ay/ /oy/ /eh/ /ah/

2.1. WFST Alignment System

Transduction in the WFST model represents all possible alignments between decoded and canonical phones sequences. In this study, 2 separate WFST alignment models are constructed for native and non-native speakers. The input and output to the WFST alignment models are the decoded phone (q_d) and the canonical phone (q_c), respectively. The WFST weights of (q_d, q_c) can be interpreted as the conditional probability of canonical phone given decoded phone, $\boldsymbol{P}(q_c|q_d)$.

2.1.1. EM Weight Training for WFST

Forward-Backward Expected Maximization (FB-EM) algorithm [8] is used to train the weights of the WFST. An initial WFST framework is constructed in such a way 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) be the pair of accented and canonical phones sequences respectively. For a given sequence pair (A_i, B_i), multiple paths through $T_{A,B}$ are possible. $T_{A,B}$'s weights are initialized such that all the phones that follows phone mapping strategy have a value of 1, otherwise the weights are floored to a significantly small value close to 0.

The FB-EM algorithm consists of 2 stages: expectation step and maximization step. In the expectation step, for each sequence pair (A_i, B_i) in the training corpus, the weight for each phone mapping is computed as follows:

1) Compute all possible alignments of (A_i, B_i) by performing compositions, $M_i = A_i \circ T_{A,B} \circ B_i, M_i \geq 1$.
2) Normalize the weights of all paths/alignments so that they sum up to 1, where the probability of a path is defined as,

$$\boldsymbol{P}(M_i) = \prod_k \boldsymbol{P}(q_{ck} \mid q_{dk}), \quad (1)$$

where k = # of phones error mappings in M_i, $i \geq 1$. The new updated $P(M_i)$ can be calculated as,

$$\boldsymbol{P}(M_i) = \frac{\boldsymbol{P}(M_i)}{\sum_i \boldsymbol{P}(M_i)}, i \geq 1. \quad (2)$$

3) For each (q_d, q_c), count instances of all error mappings as observed in all alignments M_i of all pairs of sequences (A_i, B_i). Each M_i contributes its weight to the corresponding instances of (q_d, q_c) in that alignment.

$$\boldsymbol{P}(q_{ck} \mid q_{dk}) = \sum_i N_{Mi} \boldsymbol{P}(M_i), \quad (3)$$

where $i \geq 1$, and N_{Mi} is the number of occurrences of a particular $\boldsymbol{P}(q_{ck}|q_{dk})$ in alignment M_i, this is to be done for all pairs of sequences (A_i, B_i). Probability $\boldsymbol{P}(q_{ck}|q_{dk})$ 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 $\boldsymbol{P}(q_{ck}|q_{dk})$ corresponding to each alignment and normalize them such that the total probability of all paths sum up to 1. The training iteratively uses Eqs. (1), (2), and (3) until the weights converge. At termination, the WFST weights capture the frequency of pronunciation error mappings at phone level.

2.1.2. 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 the possible alignments between decoded phones sequence and canonical phones sequence. The most likely alignment can be found as,

$$M^* = \arg\max_{Mi} \boldsymbol{P}(M_i), \quad (4)$$

and by combining Eq. (1) and (4), we get

$$M^* = \arg\max_{Mi} \prod_k \boldsymbol{P}(q_{ck} \mid q_{dk}). \quad (5)$$

The alignment between decoded and canonical phones sequences consists of the sequence of input-output labels of the WFST resulting from Eq. (5). This alignment captures error mappings at phone levels by exposing substitutions (q_{dk} ,q_{ck}), deletions (q_{dk} = *e*), and insertions (q_{ck} = *e*), where *e* represents empty phoneme. For example, the optimal alignment of "target" is shown in Fig 1 as the output from composition, $T_A \circ T_{A,B} \circ T_B$.

2.2. ME-Based Perceptual Scoring System

2.2.1. ME Features Construction and Feature Pruning

In this study, the features for the proposed MEM (Maximum Entropy Model) are pronunciation errors, i.e., substitution (q_{dk} ,q_{ck}), deletion (q_{dk} = *e*), and insertion (q_{ck} = *e*). The total number of features acquired for AE phones used in MEM training is 98. This number is reduced to 64 after eliminating the redundancies of non-error pronunciation features to a single feature through feature pruning strategy, e.g., features (aa:aa), (t:t), (d:d), (f:f), etc. are all mapped to feature (X:X).

2.2.2. ME Perceptual Modeling for Scoring

The ME modeling technique is used to learn the perceptual impact of pronunciation errors. Particularly, we wish to learn the conditional probability P(S | E) given by,

$$P(\boldsymbol{S} \mid \boldsymbol{E}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{l} \lambda_i f_i \right), \quad (6)$$

where S is the accent score, E are the pronunciation errors, l is the total # of all possible pronunciation errors at phone level and Z is the normalization factor. In order to learn the perceptual impact of different pronunciation errors, a listener evaluation of native/non-native speakers accents is conducted to collect ground truth. The human accent scores are quantized to nearest discrete score, and then the ME model is trained to predict human scores. In this manner, the ME features' weights λ_i capture the impact of perception on pronunciation errors after training. In Fig. 1, the output of the ME–Based Perceptual Scoring System shows the pronunciation errors features with its corresponding weights or penalties.

3. Experiments

3.1. Evaluation Corpus

The experimental evaluations presented in this section use the data from Native American English (N-AE) and Native Mandarin Chinese (N-MC) speakers in the CU-Accent Corpus. For training the proposed WFST alignment models, we used 24 N-MC and 55 N-AE speakers. The testing data employed for evaluation of P-WFST system, GOP, and listener evaluation consists of 13 N-MC and 5 N-AE speakers.

3.2. Listener Evaluation

For the listener evaluation, 50 N-AE listeners rated the accent scores of words in testing set. A total of 414 speech tokens were presented to 50 listeners. The procedure followed for the listener evaluation was similar to [6]. We used data from 40 N-AE listeners to train the MEM for our proposed scoring system, and data from the remaining 10 N-AE listeners as evaluation. We then computed the inter-rater correlations at speaker and word levels for the 10 N-AE listeners by following the method in [2]. These average inter-rater correlations suggest an upper bound on the level of expected correlation between human and automatic system scores.

3.3. System Training

In order to build the WFST alignment models, the decoded and canonical phones sequences of each N-AE and N-MC training data set are prepared from 12,364 and 13,654 speech tokens respectively (as explained in Sec. 2). The MEM is trained on listener evaluation data collected from 40 N-AE listeners. Here, the 40 N-AE listeners evaluated 414 speech tokens. In this study, we use the Carmel Toolkit [5] to implement FB-EM training on WFST and the composition method for alignments between decoded and canonical phones sequences, and Maxent Toolkit [9] to train the perceptual MEM from native perceptual information. After the models are obtained, the testing set is evaluated using Eq. 6 to obtain the word level accent scores.

4. Experiments and Discussions

The average correlation between human and machine accent scores is shown in Table 2. For comparison, the average inter-rater correlation is also shown. Inter-rater correlation is defined as the correlation of a rater's scores with the average scores of the rest of the listener group [2]. In order to measure the effectiveness of the GOP and P-WFST algorithms in measuring accent, two data sets are created and the correlations for the data sets are computed separately. One set (Set A) consists of N-AE and N-MC speakers while the other set (Set B) consists of N-MC only. The average inter-rater correlations at word and speaker levels for Set A are 0.73 and 0.95 respectively. From Table 2, it is observed that the proposed P-WFST system reaches a high speaker level correlation of 0.89, while attaining a word level correlation of 0.34 compared to GOP on Set A (0.89 for speaker and 0.47 for word). In Set B, the average inter-rater correlations at word and speaker levels are 0.6 and 0.81 respectively, and proposed P-WFST system attains higher correlation of 0.31 (word level) and 0.86 (speaker level) which outperforms the GOP's word and speaker level correlation at 0.27 and 0.75 respectively. At both speaker and word level, P-WFST's performance is 14.8% better than GOP's. 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 raters 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.

The next experiment investigates the relationship between number of words used to compute accent scores and the speaker level correlation performance of P-WFST and GOP. On Set A, we observe from Fig. 2 that by averaging accent scores of 4 words only, P-WST reaches higher correlation of 0.89 compared to GOP (0.87). Additionally, as the number of words increase, the algorithm performance for P-WFST and

GOP increases. On Set B, the P-WFST system reaches a higher correlation of 0.75 compared to that of GOP's (0.69) by using accent scores from 4 words. As seen in Set A, P-WFST's correlation performance increases with increase in number of words used. However, the GOP performance fluctuates as the number of words increase. It is observed that the P-WFST achieves high performance with little data (7-8 words are sufficient to provide accurate measurements). We believe that this stems from the unique approach that P-WFST applies to accent measurement, i.e., penalty assignment to pronunciation errors.

Finally, in the last experiment, word-dependent correlations are assessed for machines and human scores on Set A. From Fig. 3, we observe that the words *target*, *communication*, and *boy* exhibit high agreement for both machines and human, and therefore are the most suitable for accent assessment. On the other hand, the words *catch* and *hear* possess low correlation which reflects on both human and machine's inability to assess accent using these two words.

Table 2. *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.*

Algorithm	Correlation Coefficient			
	Word-Level		Speaker-Level	
	Set A	Set B	Set A	Set B
Inter-rater	0.73	0.6	0.95	0.81
P-WFST	0.34	0.31	0.89	0.86
GOP	0.47	0.27	0.89	0.75

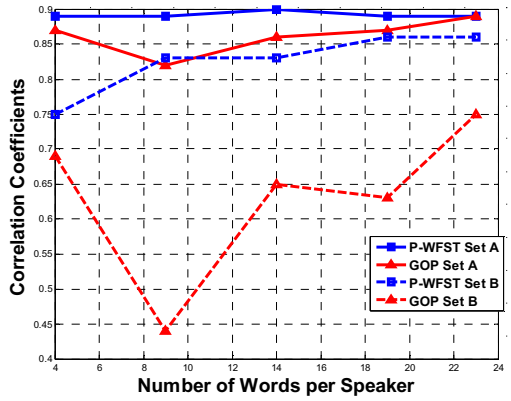

Figure 2. Variation in *speaker-level machine-human correlation with increasing number of words P-WFST and GOP performance for Set A and Set B.*

High inter-rater correlation and low human-machine correlation is observed for the words: *change* and *look*. The algorithms are less effective in assessing accent using these 2 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. We are working towards developing such a holistic approach where information from multiple sources like phones, prosody etc. is combined for accent assessment.

5. Conclusions

In this study, a new approach (P-WFST) towards accent assessment that relies on two important steps: (i) estimating pronunciation errors, and (ii) assigning perceptually motivated penalties to the pronunciation errors has been proposed. In particular, a Weighted Finite State Transducer based technique is used to detect pronunciation errors in speech. Additionally, a Maximum Entropy (ME) based technique is employed to automatically learn pronunciation error 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 outperforms 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.

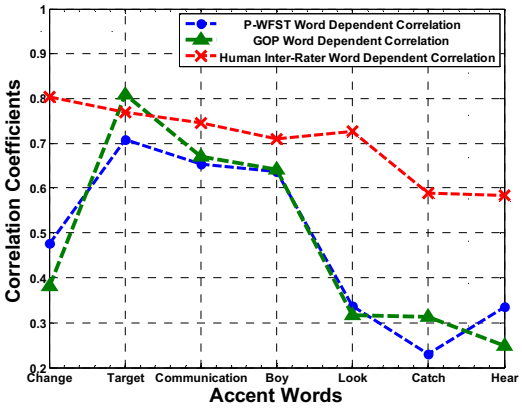

Figure 3. *Word-dependent correlation evaluated on Set A.*

6. References

[1] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic Text-Independent pronunciation scoring of foreign language student speech," in *Proc. ICSLP*, pp.1457-1460, 1996.

[2] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality", 1999.

[3] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic Pronunciation Scoring For Language Instruction," in *Proc. Int'l Conf. on Acoust., Speech and Signal Processing,* pp. 1471-1474, Munich, 1997.

[4] S. Witt, "Use of speech recognition in computer assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[5] J. Graehl. *Carmel.* Available: http://www.isi.edu/natural-language/licenses/carmel-license.html.

[6] A. Sangwan and J. H. L. Hansen, "On the use of Phonological Features for Automatic Accent Analysis," in *INTERSPEECH-2009*, pp. 172-175, 2009.

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

[8] A. P. Dempster et al., "Maximum Likelihood from Incomplete Data via the EM algorithm," Journal of the Royal Statistical Society, Series B, vol. 39, pp. 1-38, June 1977.

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