Pronunciation Assessment via a Comparison-based System

Ann Lee, James Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, Massachusetts 02139, USA
{annlee, glass}@mit.edu

Abstract

In this paper, we present preliminary results on applying a comparison-based framework to the task of pronunciation scoring. The comparison-based system works by aligning a student's utterance with a teacher's utterance via dynamic time warping (DTW). Features that describe the degree of mis-alignment are extracted from the aligned path and the distance matrix. We focus on a dataset in Levantine Arabic, a low-resource language for which there is not enough automatic speech recognition (ASR) capability available. Three different speech representations are investigated: MFCCs, Gaussian posteriorgrams, and English phoneme state posteriorgrams decoded on Levantine data. Experimental results show that the system can improve both correlation and mean squared error between machine predicted scores and human ratings compared to a template-based system.

Index Terms: pronunciation scoring, dynamic time warping, posteriorgrams

1. Introduction

The use of speech in computer-aided language learning (CALL) systems has enabled students to not only acquire vocabulary and grammatical concepts through reading but also practice pronunciation through speaking. More specifically, computer-assisted pronunciation training (CAPT) systems focus on the tasks of individual error detection and pronunciation assessment in nonnative speech [1], with the former aimed at detecting word or subword level pronunciation errors, and the latter targeted at scoring the overall fluency of an utterance. While these tasks can be further divided into processing read speech or spontaneous speech, their basic goal is the same, which is to compare a student's speech with that of a reference model.

In this paper, we focus on the task of pronunciation scoring on read speech. In early work, the reference models were stored as templates, and the student's speech was scored based on the percentage of the matching bits with that of templates [2, 3]. Later on, as automatic speech recognition (ASR) technologies improved, hidden Markov models (HMMs) were also applied to CAPT systems to model the reference speech statistically. Many of the fundamental features were based on HMM likelihood measures and posterior probability scores [4, 5, 6]. Timing scores such as phone segment duration, rate of speech and length of pauses, were also found to be highly correlated with human ratings [7, 8]. Some high-level features like recognition accuracy, confidence measures [9] and the ranking order of the correct phonemes [10] were also investigated. Another approach to model the reference speech was to build phonetic structures and use the distortion between two structures to estimate pronunciation proficiency [11, 12].

While ASR technology has its strengths, the process of building a recognizer requires a significant amount of annotated data and expertise. In addition, a new recognizer has to be built every time we want to build a CAPT system for a new target language. To address this issue, in our prior work [13], a comparison-based system was proposed for the task of mispronunciation detection in nonnative English. The system first aligns a student's utterance with a teacher's utterance via dynamic time warping (DTW). Features that describe the degree of mis-alignment are extracted from the aligned path and the distance matrix, and are then used for classifier training. The advantage of this framework is that it is language independent, and the speech representations that DTW compares can be obtained either in a fully unsupervised manner, such as Mel-frequency cepstral coefficients (MFCCs) or Gaussian posteriorgrams (GPs), or in a semi-supervised or fully supervised manner [14], such as phoneme posteriorgrams, depending on how much labeled data is available.

In this paper, we further explore this comparison-based framework in three aspects. First, we investigate the use of alignment-based features on the task of pronunciation scoring by training regressors instead of binary classifiers. Secondly, as there is no assumption about the target language for the framework, we turn our focus from nonnative English to Levantine Arabic, a low-resource language in which we do not have recognition capability. Lastly, besides MFCCs and GPs, we also explore using English phoneme state posteriorgrams decoded on Levantine data to examine the possibility of building a CAPT system for a low-resource language by taking advantage of a language with extensive resources.

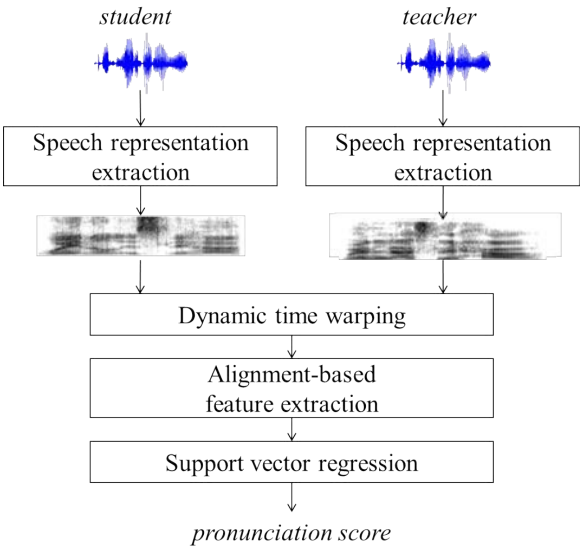


Figure 1: *System diagram. After transforming waveforms into speech representations, the system aligns the two utterances via DTW, and then extracts alignment-based features from the aligned path and the distance matrix. A support vector regressor is used for predicting an overall pronunciation score.*

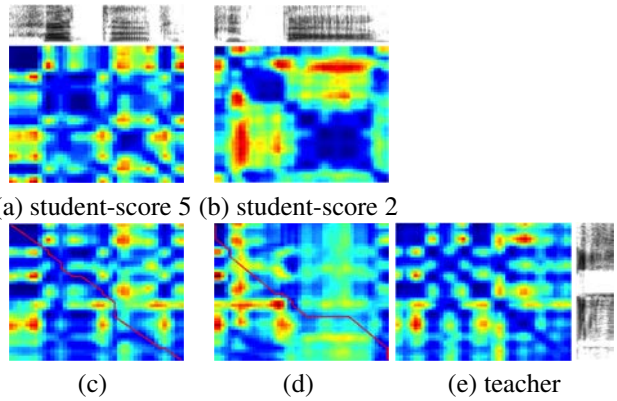
2. Corpus

The Levantine Arabic dataset consists of 21 nonnative speakers (students), including 11 males and 10 females, and 4 native speakers (teachers), including 2 males and 2 females. All students are native English speakers. Each speaker was asked to read the same 100 scripts, whose content varies from common phrases such as "Good morning" and "Thank you" to longer and more complicated sentences. Students listened to the reference audio first and then did the recording, and could repeat a recording until they were satisfied with the pronunciation. For every nonnative utterance, we have one score on a 1-5 scale for its intelligibility as decided by an expert. The scoring criterion was: 1 = many errors/unintelligible, 2 = heavy accent/difficult to understand, 3 = accented but mostly intelligible, 4 = slightly accented/intelligible, 5 = native accent/fully intelligible. There are no other human annotations on the data. After removing problematic recordings, we are left with 2064 nonnative utterances.

3. System Design

3.1. Dynamic time warping (DTW)

Fig. 1 illustrates the flowchart of the system. The first stage of the system aligns the student's utterance with a teacher's utterance through DTW. A DTW algorithm finds the optimal match between two sequences which may vary in speed. Given a teacher's utterance $T = (f_{t_1}, f_{t_2}, ..., f_{t_n})$ with n frames, and a student's utterance $S = (f_{s_1}, f_{s_2}, ..., f_{s_m})$ with m frames, an $n \times m$ distance matrix Φ_{ts} can be computed as $\Phi_{ts}(i, j) = D(f_{t_i}, f_{s_j})$, where $D(\cdot)$ denotes the distortion measure, or the distance, between two frames. DTW works by finding the path starting from $\Phi_{ts}(1, 1)$ and ending at $\Phi_{ts}(n, m)$ with the minimum accumulated distance.



(a) student-score 5 (b) student-score 2

(c) (d) (e) teacher

Figure 2: *(a) and (b) are the SSMs of two students' utterances with different scores, together with the spectrograms, and (e) is the SSM of a teacher saying the same sentence. (c) shows the alignment between (a) and the teacher, and (d) shows the alignment between (b) and the teacher. The red lines indicate the aligned paths.*

Note that the input to the DTW algorithm, i.e. f_t's or f_s's, can be of various speech representations, as long as an appropriate distortion measure can be defined. In early work, filter bank output or linear predictive features were often used [15]. More recently, posterior features have been successfully applied to facilitate not only speech recognition but also spoken keyword detection [16, 17]. The definition of a posteriorgram is as follows:

$$p_f = (P(v_1|f), P(v_2|f), ..., P(v_D|f)), \qquad (1)$$

where v_i's are the D possible models that the speech frame f might be originated from. For example, each v_i can be a single mixture in a D-component Gaussian mixture model (GMM), in which case p_f would be a Gaussian posteriorgram (GP), or each v_i can be a GMM for one single phoneme, in which case p_f would be a phoneme posteriorgram.

3.2. Alignment-based feature extraction

Fig. 2 illustrates two examples of alignments, one between a teacher's utterance and a student's utterance with a score of 5, and the other one between the same teacher's utterance and a student's utterance with a score of 2, as well as the self-similarity matrices (SSMs) of the three utterances and the corresponding spectrograms. An SSM can be obtained by aligning a sequence to itself, and thus it is symmetric on the diagonal.

We can see that a well pronounced utterance and a badly pronounced utterance have different characteristics in their alignment with the teacher. For example, for an utterance with a lower score, the aligned path would tend to be more off-diagonal, as there would be some high distortion regions along the diagonal. Also, its SSM would

be less similar to the SSM of the teacher's utterance. These observations are similar to what we had when analyzing the alignment between a reference word and a correctly pronounced word or a mispronounced word. Therefore, we can take advantage of the alignment-based features that we have designed previously. Table 1 provides an overview of each feature. More details can be found in [13].

All of the features can be extracted either on an utterance level or on a finer segmental level. In our system, we adopt an unsupervised phoneme segmentor to segment each reference utterance into smaller phoneme-like units [13]. Each distance matrix can be segmented into smaller blocks according to the segment boundaries and the aligned path. Features are extracted within each smaller unit, and we compute both the average and the standard deviation of each dimension across all the segments to form a single feature vector for an aligned pair, including the features extracted on the utterance-level.

After the alignment-based features are extracted, different regression approaches can be adopted for modeling the relationship between the features and the human ratings. In our system, we take advantage of a support vector regressor with an RBF kernel [18]. If there is more than one reference utterance for a script, we view pairs of teacher and student alignments as different instances during training, and take the average of the regressor's output for each pair during testing.

4. Experiments

4.1. Input speech representations

We explore the use of three different speech representations as inputs to our system. The first one is MFCC, for which the distance measure is defined as the Euclidean distance between two MFCC frames. The second representation is GP decoded from a 50-mixture GMM trained on all the native data (about 31 mins in total). The distance measure between two frames of GPs, p and q, can be defined as $-\log(p \cdot q)$ [16, 17].

The last representation is based on a monophone DBN-HMM English phoneme recognizer trained on the TIMIT training set to decode a set of English phoneme state posteriorgrams on the Levantine Arabic data. The DBN has 2 hidden layers (2048×2048) and a softmax layer of 183 units (3 states for each of the 61 phonemes), and takes 39-dimensional MFCCs stacked with 10 neighboring frames as input. As a result, each frame of the English phoneme state posteriorgrams is a 183-dimensional vector, and the distance measure can be also defined as the inner product distance.

Note that the first two speech representations can be obtained in a fully unsupervised manner. Though the last speech representation requires a carefully transcribed corpus in English, it does not require any phonetic labels in Levantine Arabic, a language with relatively few resources available.

Table 1: *The alignment-based features*

Aligned path & diagonal	
acc_path	accumulated distance along the aligned path
avg_path	*acc_path* normalized by path length
std_path	standard deviation of the distance along the aligned path
acc_diag	accumulated distance along the diagonal
avg_diag	*acc_diag* normalized by diagonal length
std_diag	standard deviation of the distance along the diagonal
diff_acc_p_d	*acc_path* − *acc_diag*
diff_avg_p_d	*avg_path* − *avg_diag*
ratio_avg_p_d	*avg_path* / *avg_diag*
max_seg_ratio	the length of the longest horizontal or vertical segment / path length
Distance matrix ($disMat$)	
avg_block	average distance within the block
std_block	standard deviation of the distance within the block
Duration	
dur_ratio	ratio between the length of the two sequences
diff_rel_dur	difference between the length of the two sequences that are normalized by the length of each full utterance
ratio_rel_dur	ratio between the length of the two sequences that are normalized by the length of each full utterance
Comparison with the reference	
diff_avg_block	*avg_block* − the average of the corresponding block in $SSM_{teacher}$
diff_avg_p_t	*avg_path* − the aligned path in the corresponding block in $SSM_{teacher}$
diff_avg_d_t	*avg_diag* − the aligned path in the corresponding block in $SSM_{teacher}$
diff_mat_t	element-wise difference between the rewarped $disMat$ and $SSM_{teacher}$
diff_s_t	element-wise difference between $SSM_{student}$ and $SSM_{teacher}$
hog_diff_mat_t	difference between the histograms of oriented gradients of the rewarped $disMat$ and $SSM_{teacher}$ [19, 20]
hog_diff_s_t	difference between the histograms of oriented gradients of $SSM_{student}$ and $SSM_{teacher}$

4.2. Experimental setup

We take advantage of the same English phoneme recognizer to first remove the silences at the beginning and the end of each utterance. Then, all waveforms are trans-

formed into 39-dimensional MFCCs every 10-ms, including first and second order derivatives, for the following GPs or phoneme state posteriorgrams decoding.

As there is no phonetic transcription for the data and thus we do not have recognition capability in Levantine Arabic, the baseline simulates a template-based system that scores an utterance based only on *acc_path*, *avg_path* and *std_path*. For evaluation, we run 100 iterations of 5-fold speaker-level cross validation using data from all 21 speakers. Only alignments between speakers with the same gender are considered. We compute both Pearson's correlation and the mean squared error (MSE) between the machine predicted scores and the human ratings.

4.3. Results

Experimental results are shown in Table 2. For all three speech representations, the comparison-based system obtains improvements relative to the template-based baseline in a range of 4.5% to 11.6% in correlation, and 3.8% to 15.9% in MSE. These results imply that the shape of the aligned path or the appearance of the distance matrix can provide more information about the quality of the pronunciation than alignment scores can do. These findings also agree with the findings we had in the task of mispronunciation detection. Using features extracted on the utterance level produces better results in both correlation and MSE than using features extracted on the phone level. A possible explanation is that aggregating the errors, i.e. the degree of mis-alignment, is better than averaging them. However, unlike our previous findings, there is no clear conclusion as to whether combining features from both levels can really achieve better performance.

Among the three speech representations, English phoneme state posteriorgrams gives the best result and also the largest improvement. This improvement most likely comes from the human supervision involved during English recognizer training for decoding posteriorgrams. The discriminative training process helps reduce mis-alignments from difference between speaker characteristics. Nevertheless, the high performance of the English phoneme state posteriorgrams suggests that high-resource language resources can be leveraged for training recognition on low resource languages in the context of a comparison-based approach. Because the alignment-based feature extraction process can be made independent from speech representation, a comparison-based approach can be feasibly integrated with the use of high-resource languages as training data.

4.4. Discussion

To further investigate how each type of alignment-based feature contributes to the task of pronunciation scoring, we focus on the English phoneme state posteriorgrams and repeat the 5-fold speaker-level cross validation by training on one single feature (extracted on both utterance-level and phone-level) at a time. Fig. 3 shows the correlation between the system output and human ratings for each feature.

Table 2: *Correlation and mean squared error between the machine predicted scores and the human ratings under different settings*

	MFCC	GP	English phoneme state posteriorgrams
Correlation			
Baseline	0.492	0.507	0.510
Utterance-level	0.526	0.536	0.559
Phone-level	0.511	0.526	0.534
Full system	0.523	0.535	**0.569**
Mean squared error			
Baseline	0.543	0.539	0.542
Utterance-level	0.513	0.509	0.491
Phone-level	0.519	0.516	0.508
Full system	0.522	0.507	**0.456**

First, note that the overall system performance is better than the results from using any single feature alone. This agrees with the results from several previous studies [6] which found that combining different scoring features can compensate for the weakness of each and produce a score that better correlates with human ratings.

Among the four different feature categories, the last one which compares the aligned path or the distance matrix with the self aligned path or the teacher's SSM obtains the best results on average. This could explain part of the reason why the comparison-based system can improve upon template-based approaches. Because the SSM from the teacher represents an optimal match, comparing it against the distance matrix can indicate proximity to a perfect match in a way that is different from template-based approaches relying only on alignment scores.

Moreover, a system based on *acc_path* or *acc_diag* performs better than a system based on *avg_path* or *avg_diag*. This again indicates that averaging or normalizing with respect to length may dampen the effect of high distortion regions. In line with previous work [7] indicating that utterance length is highly correlated with human ratings, the accumulated scores which have such information embedded also correlate better with human ratings. Although there is a chance that students may cheat the system by reading very quickly, there did not appear to be students circumventing the system in this way in our dataset.

5. Conclusion and Future Work

In this paper, we have explored the use of a comparison-based system in the task of pronunciation scoring. Experimental results have shown that, as in the task of mispronunciation detection, adopting alignment-based features that are extracted from the aligned path and the distance matrix can also improve system performance in predict-

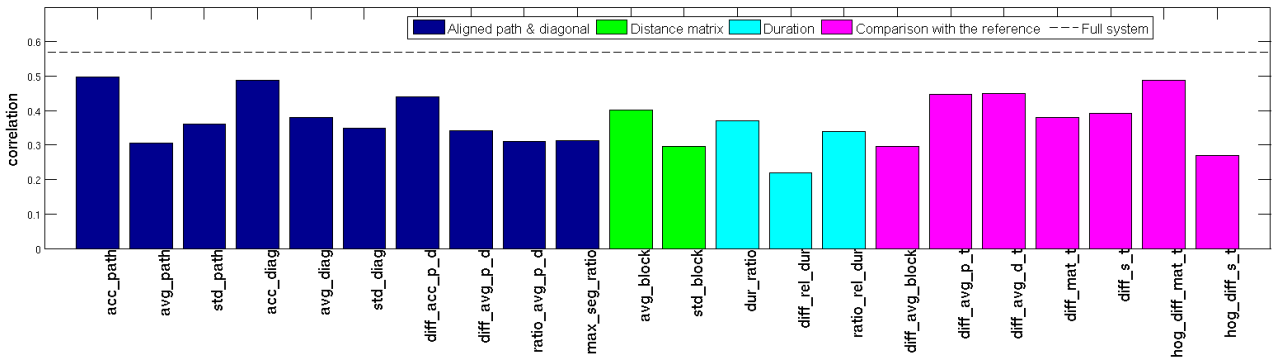


Figure 3: *Correlation between system output scores and human ratings based on a single feature*

ing pronunciation scores. The comparison-based system can be viewed as a combination of template-matching and classifier-based approaches. In fact, many of the alignment-based features are similar to ASR-based features that have been proved useful in pronunciation scoring. For example, comparing the structure of student and teacher SSMs is in some sense similar to comparing their phonetic structures [11, 12]. Features involving time comparisons might also reflect underlying durations of phoneme-like units.

Because the dataset we have collected is an initial attempt at gathering nonnative speech in a low-resource language, our current experiments are based on a relatively small dataset compared to that of previous work. As efforts continue to gather more data, we intend to examine system performance on larger-scale datasets, with the hope of enhanced performance due to greater amounts of training data. Running experiments on a dataset whose size is comparable with those in other studies can also allow us to have a fair comparison between absolute system performance. Future work should explore training the regressor from alignments in one language and testing on the other language to see whether misalignment patterns may be universal, and experimenting with speech representations that are more robust to different channel characteristics so that we can leverage more data from different sources.

6. Acknowledgements

The authors would like to thank Wade Shen for providing the dataset, and Ekapol Chuangsuwanich, Yu Zhang, Yaodong Zhang and Hung-An Chang for their help with the DBN-HMM recognizer.

7. References

[1] Eskenazi, M., "An overview of spoken language technology for education", in Speech Communication, 2009.

[2] Kewley-Port, D., Watson, C., Maki D. and Reed D., "Acoustic-articulatory inversion", in proc. ICASSP, 1987.

[3] Wohlert, H., "Voice input/output speech technologies for German language learning", in Die Unterrichtspraxis/Teaching German, 1984.

[4] Witt, S. M. and Young, S. J., "Phone-level pronunciation scoring and assessment for interactive language learning", in Speech Communication, 2000.

[5] Neumeyer, L., Franco, H., Digalakis, V. and Weintraub, M., "Automatic scoring of pronunciation quality", in Speech Communication, 2000.

[6] Franco, H., Neumeyer, L., Digalakis, V. and Romen, O., "Combination of machine scores for automatic grading of pronunciation quality", in Speech Communication, 2000.

[7] Cucchiarini, C., Strik, H. and Boves, L., "Automatic evaluation of Dutch pronunciation by using speech recognition technology", in proc. ASRU, 1997.

[8] Bernstein, J., De Jong, J., Pisoni, D. and Townshend, B., "Two experiments on automatic scoring of spoken language proficiency", in proc. Integrating Speech Technology in Learning, 2000.

[9] Cincarek, T., Gruhn, R., Hacker, C., Noth, E. and Nakamura, S., "Automatic pronunciation scoring of words and sentences independent from the non-native's first language", in Computer Speech and Language, 2009.

[10] Chen, J.-C., Jang, J.-S., Li, J.-Y. and Wu, M.-C., "Automatic pronunciation assessment for Mandarin Chinese", in proc. ICME, 2004.

[11] Minematsu, N., "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances", in proc. ICSLP, 2004.

[12] Suzuki, M. Dean, L., Minematsu, N. and Hirose, K., "Improved structure-based automatic estimation of pronunciation proficiency", in proc. SLaTE, 2009.

[13] Lee, A. and Glass, J., "A comparison-based approach to mispronunciation detection", in proc. SLT, 2012.

[14] Lee, A., Zhang, Y. and Glass, J., "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams", in proc. ICASSP, 2013.

[15] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", in IEEE Trans. on Acoustics, Speech and Signal Processing, 1978.

[16] Hazen, T. J., Shen, W. and White, C., "Query-by-example spoken term detection using phonetic posteriorgram templates', in proc. ASRU, 2009.

[17] Zhang, Y. and Glass. J. R., "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams", in proc. ASRU, 2009.

[18] Chang, C.-C. and LIN, C.-J., "LIBSVM: A library for support vector machines", in ACM Transactions on Intelligent Systems and Technology, 2011. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[19] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection", in CVPR, 2005.

[20] Muscariello, A., Gravier, G. and Bimbot, F., "Towards robust word discovery by self-similarity matrix comparison", in proc. ICASSP, 2011.