AUTOMATIC ASSESSMENT OF ENGLISH LEARNER PRONUNCIATION USING DISCRIMINATIVE CLASSIFIERS

Mauro Nicolao, Amy V. Beeston, Thomas Hain

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK

ABSTRACT

This paper presents a novel system for automatic assessment of pronunciation quality of English learner speech, based on deep neural network (DNN) features and phoneme specific discriminative classifiers. DNNs trained on a large corpus of native and non-native learner speech are used to extract phoneme posterior probabilities. A part of the corpus includes per phone teacher annotations, which allows training of two Gaussian Mixture Models (GMM), representing correct pronunciations and typical error patterns. The likelihood ratio is then obtained for each observed phone. Several models were evaluated on a large corpus of English-learning students, with a variety of skill levels, and aged 13 upwards. The cross-correlation of the best system and average human annotator reference scores is 0.72, with miss and false alarm rate around 19%. Automatic assessment is 81.6% correct with a high degree of confidence. The new approach significantly outperforms spectral distance based baseline systems.

Index Terms— Pronunciation assessment, Computer-Assisted Language Learning, DNN-GMM, binary classifier.

1. INTRODUCTION

As automatic assessment tools permeate teaching methodologies, reliable automatic assessment of English learner speech is of increasing interest. Interactive language learning tools incorporate a variety of approaches, but assessment of pronunciation quality remains a particularly challenging task, especially with adolescent students. Approaches to computer-assisted pronunciation training were reviewed in [1, 2], and use a variety of metrics designed to assess pronunciation at the fine-grained phonetic level [3]. Nonetheless, studies in this area tend to score learners' speech on longer time-intervals, reporting assessments in units of words [4, 5] or longer, e.g. per-sentence or per-student measures as in [6].

This paper describes a method for phone-level pronunciation error detection developed as part of a research project on language learning for Dutch learners of English. The project has collected a large corpus of classroom recordings which is described in more detail below. Our approach is similar to that of [4, 5] in that a student's attempt to pronounce a given phrase is compared directly against an audio example provided by their teacher.

Our proposed method produces a phoneme-level assessment, which is a far more challenging task than word, sentence or student level assessment. We introduce a binary error classification regime that allows an efficient pronunciation assessment, and which benefits from advanced acoustic modelling with deep neural networks (DNNs). The new method is also computationally efficient, as it uses DNN features and GMM-based binary classification rather than automatic speech recognition, as recent related work describes [7]. The cross-correlation of our best system and average human annotator reference scores is 0.72, with miss and false alarm rate around 19%. Automatic assessment is 81.6% correct with a high degree of confidence.

2. PRONUNCIATION ASSESSMENT

The objective is to assess the pronunciation quality of an unknown student utterance, U_s. It is assumed that a reference utterance U_t exists, spoken by a teacher. The system then outputs a vector $\mathbf{S} = \{s_i\}$ with a pronunciation quality assessment score for each of the M phonemes in U_s. By design, teacher and student utterances have the same word content, $\mathbf{W}_s = \mathbf{W}_t$. Moreover, a proficient learner should exactly match their teacher's phonetic sequence, $\mathbf{P}_s = \mathbf{P}_t$.

Given two recordings of the same text prompt – one from the student, $\mathbf{O}_s$, and one from the teacher, $\mathbf{O}_t$ – we would like to compute the probability that the learner recording mimics the pronunciation in the teacher reference. Alternatively one can ask if the teacher's reference is a good predictor for the student utterance, i.e. we would like to compute $P(\mathbf{O}_s|\mathbf{O}_t)$.

With the reference $\mathbf{O}_t$, the word $\mathbf{W}_t$ and the phonetic $\mathbf{P}_t$ sequences are also determined. Hence, $P(\mathbf{O}_s|\mathbf{O}_t)$ can be written as

$$P(\mathbf{O}_s|\mathbf{O}_t) \quad = \quad \frac{P(\mathbf{O}_s, \mathbf{W}_t, \mathbf{P}_t, \mathbf{O}_t)}{P(\mathbf{W}_t, \mathbf{P}_t, \mathbf{O}_t)} \qquad (1)$$

For estimating the above we use the fact that teacher and student words are the same, i.e $\mathbf{W}_t = \mathbf{W}_s = \mathbf{W}$, but also assume that the two phonetic sequences are identical, $\mathbf{P}_t = \mathbf{P}_s = \mathbf{P}$. As $P(\mathbf{P}, \mathbf{W}, \mathbf{O}_t)$ depends only on the reference

This work was undertaken in collaboration with ITSLanguage BV (http://www.itslanguage.nl)

Set name	Use	Acquisition	# talkers	Age	# files	# hours
$\text{INA}_{\text{ph 1–2}}$	GMM training	Apr 2013	238	13+	6,252	3.05
$\text{INA}_{\text{ph 3–6}}$	GMM test and regression tree training	Apr 2013	222	13+	6,640	2.99
INT	DNN and acoustic model training	May 2013	598	13+	88,697	46.49
INY	Pronunciation reference (U_t)	Dec 2013	8	23+	1,869	1.68

Table 1. ITSLanguage data subsets. Annotated learner recordings in INA are split into two portions of around 3 hours duration each. Learner material in INT is used in training, and INY contains teacher recordings used as pronunciation references.

segmentation, it is constant for every similar utterance of the learner and (1) can be approximated as

$$\frac{P(\mathbf{O}_s, \mathbf{W}_t, \mathbf{P}_t, \mathbf{O}_t)}{P(\mathbf{W}_t, \mathbf{P}_t, \mathbf{O}_t)} \propto P(\mathbf{P})P(\mathbf{O}_s, \mathbf{O}_t|\mathbf{P}). \quad (2)$$

The above phoneme sequence prior is constant for all observations. Assuming segmentation information for phonemes, $\mathbf{P} = \{r_i\}$, both student and teacher feature sequences can be split into the phone-related sets:

$$P(\mathbf{O}_s, \mathbf{O}_t|\mathbf{P}) = \prod_{i=1}^{M} P(\mathbf{O}_s^i, \mathbf{O}_t^i|r^i) \quad (3)$$

The paired sets, $\mathbf{O}_s^i$ and $\mathbf{O}_t^i$, are normally of different length. In order to give each realisation of the phone the same importance, the duration is normalised to a fixed length L. The simplest solution might set $L = 1$ and chose, for example, either the central feature vector, $\mathbf{O}^i = \mathbf{o}^{i,central}$ or an average of the feature vectors within the phone r_i along each dimension. If the feature domain is assumed to be continuous along each dimension (such as in the posterior features domain), another solution would be to resample or interpolate the values on each dimension to achieve feature sets with same length L, e.g. $\mathbf{O}_s^{i,L}$ and $\mathbf{O}_t^{i,L}$ with $L = 20$. Each element of the product in (3) can be thus modelled as a phoneme-dependant mixture of Gaussians (GMM_i).

The problem then can be turned into a binary classification problem, where phoneme-level scores for each student utterance $\mathbf{O}_s$ are computed. For pairwise aligned data, $\mathbf{O}^{i,L} = [\mathbf{O}_s^{i,L}, \mathbf{O}_t^{i,L}]$, each student's phone is judged to be well pronounced (C=correct) when $P(\text{C=correct}|\mathbf{O}^{i,L}) > P(\text{C=error}|\mathbf{O}^{i,L})$. This is equivalent to

$$\frac{P(\mathbf{O}^{i,L}|\text{C=correct})}{P(\mathbf{O}^{i,L}|\text{C=error})} > \frac{P(\text{C=correct})}{1 - P(\text{C=correct})} \quad (4)$$

where the right hand side of (4) serves as a threshold T, which depends on the degree of proficiency of the learners.

3. THE ITSLANGUAGE DATASET

The dataset underpinning this work comprises recordings of native-speakers and learners of English, a portion of which has mispronunciation annotated at a phonetic level.

Comparison	vs.	Agreement	CC	vs.	Agreement	CC
a1	a2	0.858	0.434	**R**	0.947	0.816
a2	a3	0.782	0.412	**R**	0.911	0.649
a3	a1	0.818	0.523	**R**	0.871	0.678

Table 2. Inter-annotator analysis. *Left:* pairwise comparisons showing Agreement (where annotators scores match) and cross-correlation (CC, which detects similar behaviour across sequences of values) [3]. *Right*: Each annotator is compared with the combined reference, $\mathbf{R}$.

Recordings were made via an online learning environment, by mainly Dutch pupils in schools across the Netherlands. Working individually, the learners read items from an ordered list of 193 text prompts (both words and short sentences), re-recording each item until satisfied with their pronunciation. Recorded with headset microphones, audio signals were stored in MS-WAVE format (22.05 kHz, 16-bit). Many students performed the recording task simultaneously; a high degree of background noise was therefore present in each classroom. Metadata for each learner detailed their age, mother tongue and other languages spoken. Further, their familiarity with English was quantified in several ways: the number of years learning, a self-reported confidence score, and the Common European Framework level of the class in which the student was enrolled. The learner dataset consisted of around 80 hours of raw speech. Recordings with prominent distortion were removed using clipping detection. Items with partial, missing, or inappropriate speech content were filtered out by aligning the audio with the known text prompt (using a British English acoustic model and multiple-pronunciation dictionary). Table 1 illustrates learner subsets selected for the current work: around 6 hours (INA_{ph}) was selected for phone-level annotation and c. 46.5 hours (INT) was used in system development as detailed below.

Pronunciation references (INY) for the 193 test items were recorded by proficient British English speakers. A high quality microphone and quiet room were used, but other conditions replicated those described above. Teacher data comprised recordings by 4 native-speaking adults (2 male, 2 female) and 4 non-native adults (2 male, 2 female, with Dutch, Austrian, and Flemish backgrounds).

The annotation dataset (INA) included utterances from as

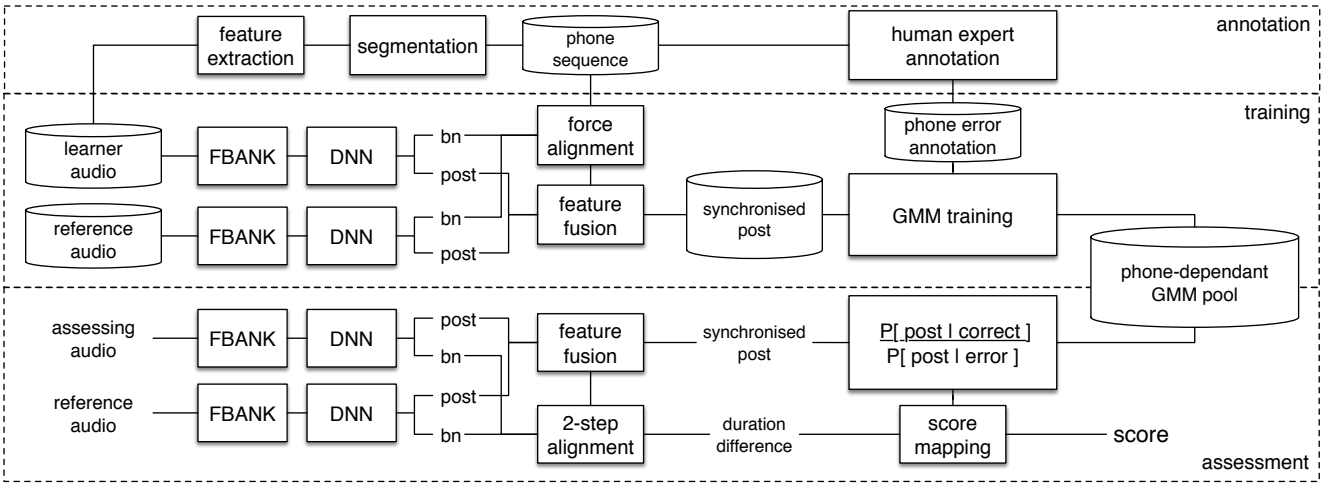

Fig. 1. Pronunciation evaluation framework showing stages of annotation (*top*), training (*middle*) and assessment (*bottom*).

many learners as possible, balancing gender, age and learner level. Around half was annotated at the phone-level (cf. INA_{ph} in Table 1), and half at the word-level (unused in the present study). These utterances were aligned (allowing for multiple pronunciations) in order to provide a standard phone sequence reference, $\mathbf{P}_a$, which was presented alongside the audiofile in an online annotation tool. Three native Dutch-speaking phoneticians (a1, a2, a3) acted in the role of teacher to assess the correctness of the learners' pronunciation. Table 2 quantifies inter-annotator consistency. To handle the frequent differences of opinion, a combined reference, $\mathbf{R}$, comprised the average of the three responses for each phone in each utterance. This located regions that all three annotators considered well-pronounced or mispronounced, and ambiguous regions causing disagreement. This approach differs from [5], where items that were not unanimously labelled were removed. Moreover, it reflects the fact that an 'error' resists clear definition since pronunciation varies continuously between native and unintelligible [2]. A threshold on $\mathbf{R}$ then highlights consistently well- or poorly-spoken phones.

4. EVALUATION FRAMEWORK

The method described in Section 2 was implemented with the architecture outlined in Figure 1. The observation vector in (4) has no restriction on the type of features that can be used. Here, posterior probability features (labelled *post*) were extracted, using a deep neural network (DNN). Two DNNs were tested: DNN_{UK}, a 2-layer network trained on clean British English pronunciation using the WSJCAM0 corpus [8], and $\text{DNN}_{\text{UK+INT}}$, a 4-layer network trained using both British and Dutch-accented student audio (WSJCAM0 and INT). As input, both DNNs used a 15-frame span vector with 23 filterbank coefficients per frame. The bottleneck layer had 26 coefficients and the output layer had 144 monophone states (English phone set). Posterior features were extracted from both the $\text{INA}_{\text{ph 1–2}}$ learner corpus and one or more teachers in INY. The three *post* values measured for each phone state were combined into a single value. The output of the bottleneck layer (*bn*) was also used to train the triphone-based GMM-HMM acoustic model used in alignment.

During training, learner and teacher audio fragments were fused in a pairwise manner, phone-by-phone, by means of the annotation sequences $\mathbf{P}_a$. During assessment, a '2-step' process first obtained a target phone sequence from the teacher's audio using a multiple-pronunciation dictionary, and secondly used this sequence in forced alignment of the student's recording. Teacher and learner phones typically have different durations; same-length feature sets were created here by interpolation of the extracted vectors along each dimension, setting length $L = 20$. The alignment and fusion processes thereby generated the phone-comparison vectors $\mathbf{O}^{i,L} = [\mathbf{O}_s^{i,L}, \mathbf{O}_t^{i,L}]$ of (4). In training, these vectors were then grouped into 47 (one per phone i), giving sets of $L \times N_i$ vectors, where N_i is the number of phone realisations. Each set was further split using the reference $\mathbf{R}$ with a threshold at 0.5, which provided a binary decision label based on the annotators' majority opinion. After this, $\{\mathbf{O}^{i,L}\}_{N_i}^{C}$ stored all the correct phone pronunciations, and the mispronunciations were gathered in $\{\mathbf{O}^{i,L}\}_{N_i}^{E}$. These sets were input to the GMM expectation-maximisation training, allowing computation of $P(\mathbf{O}^{i,L}|\text{C=correct})$ and $P(\mathbf{O}^{i,L}|\text{C=error})$. In total, 94 GMM_i were created, each comprising 64 Gaussian functions.

During assessment, the same feature extraction process took place, using the 2-step alignment discussed above. The left side of (4) was then computed using the synchronised feature set, resulting in a wide range of continuous likelihood ratio values. These were mapped into the [0,1] interval using a regression tree which was optimised using the annotated data of the GMM training, a measure of the time-discrepancy

SYSTEM	U_t (gender)	U_s	MAPPING	FAR	MISS	CC	ACC	F-SCORE	NCE
Baseline [D+E+S+Δ]	INY51 (F)	$\text{INA}_{\text{ph 1–2}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.443	0.365	0.443	0.695	-	0.008
DNN_{UK}+INY51	INY51 (F)	$\text{INA}_{\text{ph 1–2}}$	none	0.239	0.207	0.426	0.763	0.327	N.A.
DNN_{UK}+INY51	INY51 (F)	$\text{INA}_{\text{ph 1–2}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.184	0.193	0.716	0.816	0.388	0.397
DNN_{UK}+INY51	INY51 (F)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.252	0.264	0.613	0.747	0.261	0.447
DNN_{UK}+INY51+D	INY51 (F)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.252	0.262	0.614	0.748	0.262	0.446
DNN_{UK}+INY51+D+Δ	INY51 (F)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.258	0.255	0.616	0.742	0.260	0.451
$\text{DNN}_{\text{UK+INT}}$+INY51+D+$\Delta$	INY51 (F)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.297	0.275	0.581	0.704	0.229	0.559
$\text{DNN}_{\text{UK+INT}}$+MULTI+D+$\Delta$	INY51 (F)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.288	0.284	0.576	0.712	0.232	0.473
DNN_{UK}+INY51+D+Δ	INY52 (M)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.250	0.241	0.574	0.751	0.504	0.371
$\text{DNN}_{\text{UK+INT}}$+INY51+D+$\Delta$	INY52 (M)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.291	0.281	0.582	0.710	0.231	0.558
$\text{DNN}_{\text{UK+INT}}$+MULTI+D+$\Delta$	INY52 (M)	$\text{INA}_{\text{ph 3–6}}$	reg.tree($\text{INA}_{\text{ph 1–2}}$)	0.278	0.277	0.587	0.722	0.240	0.470

Table 3. Comparison of phone-level mispronunciation detection scores predicted by the system, $\mathbf{S}$, and marked in the human annotators' reference, $\mathbf{R}$. *Top:* baseline system. *Middle:* system development using a matched teacher voice for training and assessment stages of the framework. *Bottom:* Comparisons using a mis-matched teacher voice.

between teacher and student phone durations (D), and their differential values (Δ). A pronunciation error resulted when mapped scores fell beneath a threshold whose level sets the system's strictness. Here, the rate of undetected mispronunciations (MISS) balanced the well-pronounced phones marked as errors (false alarm rate, FAR).

5. EXPERIMENTS

The evaluation framework was used to assess learner pronunciation in the (unseen) $\text{INA}_{\text{ph 3–6}}$ dataset. For this, the system outcome, $\mathbf{S}$, was compared with the combined human reference, $\mathbf{R}$, using six common information retrieval indices. Of these, CC, FAR and MISS were previously defined; additionally, Accuracy (ACC), F-score, and normalised cross-entropy (NCE) are displayed in Table 3. ACC gives the percentage of phones that were correctly assessed. The F-score combines recall and precision rates [5]. NCE concerns the mutual information between the correctness of the mispronunciation detection and the confidence score in making that decision [9]. For the perfect system, FAR and MISS values are low, and CC, ACC, F-score and NCE are high.

The baseline system reported in [10] performed pairwise comparison of temporal and spectral acoustic features averaged at the phone-level (duration, D, energy, E, and spectral shape, S). These were combined into a single score in the [0,1] range using a regression tree as described above.

Table 3 shows the evaluation framework's improvement (*middle*) over the baseline (*top*), and outlines the relative gains (or losses) arising as each component is introduced. Annotated learner speech material (U_s) is compared against utterances recorded by a single teacher (here, U_t=INY51, female), simulating the manner in which this learning environment would be used in the classroom. The proposed system uses audio data from the selected teacher to train the GMMs. System performance is loosely similar to that of the baseline, even without the regression tree mapping (e.g., CC is 0.426 and 0.443 respectively). The performance is boosted substantially when the regression tree is tuned on student utterances used in the GMM training stage (CC=0.716 for $\text{INA}_{\text{ph 1–2}}$). Importantly, cross-correlation remains high (CC=0.613) when this system is then tested on *unseen* learner audio, i.e. when $\text{INA}_{\text{ph 3–6}}$ is used for U_s, simulating a new student group. A slight improvement is observed by introducing D and Δ as input to the regression tree (CC=0.616). However, inclusion of the student data in the DNN creation ($\text{DNN}_{\text{UK+INT}}$) brings about a small reduction in performance (CC=0.581). This arises from the higher diversity of pronunciation captured in the DNN posterior distributions.

An ideal system would allow introduction of a new teacher without penalty (e.g., U_t=INY52, male, in Table 3, *bottom*). However, a drop in performance occurred when the DNN_{UK}+INY51+D+Δ system was used with INY52. To lessen the dependency on the training teacher, the GMM was created using multiple teacher references (MULTI). When a teacher is involved in both test and training (i.e., U_t=INY51, female), there is no further benefit from this MULTI condition. However, a different pattern of results emerges for the mis-matched teacher condition using U_t=INY52, male, where the MULTI condition gives an improvement in system performance (CC increases from 0.574 to 0.587).

6. CONCLUSIONS

This paper introduced a novel system for automatic assessment of pronunciation quality of English learner speech, based on DNN features and phoneme-specific discriminative classifiers. In this challenging phone-level decision task, one in which even expert annotators often disagree, the proposed method achieved good accuracy values with high decision confidence. Improvements in DNNs and annotation quality might address the low performance in FAR and MISS.

7. REFERENCES

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, Oct. 2009.

[2] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *IS ADEPT*, Stockholm, SE, June 2012, pp. 1–8.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, Feb. 2000.

[4] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *SLT 2012*, Miami, FL, Dec. 2012, pp. 382–387.

[5] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via Dynamic Time Warping on Deep Belief Network-based posteriorgrams," in *ICASSP 2013*, Vancouver, BC, May 2013, pp. 8227–8231.

[6] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, Mar. 2008.

[7] A. Metallinou and J. Cheng, "Using Deep Neural Networks to improve proficiency assessment for children english language learners," in *INTERSPEECH 2014*, Singapore, Sept. 2014, pp. 1468–1472.

[8] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *ICASSP 1995*, 1995, pp. 81–84.

[9] W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. J. Brady, "Estimating and Evaluating Confidence for Forensic Speaker Recognition," in *ICASSP 2005*, Philadelphia, PA, 2005, pp. 717–720.

[10] A. V. Beeston, M. Nicolao, and T. Hain, "Pairwise audio comparison for visualisation of mispronunciation," in *2nd UK Speech Conference*, Cambridge, UK, Sept. 2013.