PERFORMANCE IMPROVEMENT OF AUTOMATIC PRONUNCIATION ASSESSMENT IN A NOISY CLASSROOM

†*Yi Luan,* †*Masayuki Suzuki,* ‡*Yutaka Yamauchi,* †*Nobuaki Minematsu,* †*Shuhei Kato,* †*Keikichi Hirose*

†The University of Tokyo, ‡Tokyo International University
†{luanyi, suzuki, mine, kato, hirose}@gavo.t.u-tokyo.ac.jp, ‡yyama@tiu.ac.jp

ABSTRACT

In recent years Computer-Assisted Language Learning (CALL) systems have been widely used in foreign language education. Some systems use automatic speech recognition (ASR) technologies to detect pronunciation errors and estimate the proficiency level of individual students. When speech recording is done in a CALL classroom, however, utterances of a student are always recorded with those of the others in the same class. The latter utterances are just background noise, and the performance of automatic pronunciation assessment is degraded especially when a student is surrounded with very active students. To solve this problem, we apply a noise reduction technique, Stereo-based Piecewise Linear Compensation for Environments (SPLICE), and the compensated feature sequences are input to a Goodness Of Pronunciation (GOP) assessment system. Results show that SPLICE-based noise reduction works very well as a means to improve the assessment performance in a noisy classroom.

Index Terms— pronunciation evaluation, noise reduction, GOP , SPLICE , CALL

1. INTRODUCTION

Compared to traditional language education methodologies, CALL systems have many potential benefits [1]. CALL systems are faster and cheaper, which allow learners to get feedback immediately and study by themselves without requiring the sole attention of a teacher. In CALL systems, a good pronunciation evaluation method is needed to inform learners about their proficiency and to correct their pronunciations. However the evaluation methods in current CALL systems are still not as good as human teachers. Previous research has shown that computer evaluation systems are less robust than human teachers when facing poor quality audio files, while human evaluation remains consistent [2]. Many factors may affect the quality of a audio file, including using low quality microphones, setting up recording software incorrectly, and background noise generated from air conditioners, other learners, etc. Recently, more and more educational facilities have begun to utilize CALL systems during classes. When ASR-based technologies are used, utterances from other learners will be recorded at the same time, which will negatively impact the performance of automatic evaluation approaches, especially when surrounding students are very active.

In order to improve the robustness of automatic pronunciation evaluation in CALL systems, we investigate the effect of using a noise reduction technique in automatic pronunciation proficiency estimation. Here, we test SPLICE [3], which is a noise reduction algorithm in the presence of additive noise, channel distortion, or a combination of the two. SPLICE is efficient especially when the distortion characteristics are known beforehand and is used in ASR systems to reduce the degradation caused by mismatches between training data and testing environments. In a noisy classroom, the main noise sources are speech from surrounding students and some microphone noise caused by touching a microphone to adjust its position and direction. Therefore it seems reasonable to use SPLICE to solve the problem.

In this paper we use a GOP-based pronunciation scoring system [4] as baseline system and evaluate the effect of SPLICE on it. GOP is an acoustic likelihood-based method for automatic pronunciation assessment based on Hidden Markov Models (HMMs). GOP is especially efficient in evaluating the proficiency of pronunciations [5]. As well as for read speech, GOP was also used to evaluate the utterances recorded in shadowing practices [6]. We conduct two sets of experiments to evaluate the method. The first set utilizes the English Read by Japanese (ERJ) [7] database which contains recordings of English read by Japanese students and human pronunciation scores for each recording. The second set consists of real data from foreign language learners which was recorded via a shadowing exercise. Also in this database, the proficiency score is provided to each utterance by human teachers. For each database, we calculated the correlation between the human scores and the GOP scores of the recorded utterances to compare the results. Both of the experiments show the effect of SPLICE on improving the correlation between human scores and machine scores. The result of the ERJ experiment shows a relative correlation increase of 13.4% on average. The experiment based on real data shows a relative average correlation increase of 6.75%.

The remainder of this paper is organized as follows. In

Section 2, we give a brief review of the basic SPLICE algorithm. In section 3, the basic GOP algorithm is introduced. The design and results of the experiments are described in Section 4. Section 5 provides analysis, discussion and future work and concludes the paper.

2. AN OVERVIEW OF SPLICE

SPLICE is a noise reduction method used in ASR to remove consistent degradation of speech cepstra. SPLICE is not constrained to any specific kind of noise in the sense that it does not model a specific kind of noise, but models the transformation probabilistically from noisy speech to its clean version. It does not include any assumptions about how the noise is produced and thus can be used to model any combination of additive noise or convolutional channel.

Generally speaking, the transformation of clean speech to noisy speech is nonlinear in the cepstrum domain. Therefore, in order to train SPLICE, first separate the space of noisy features into several subspaces according to a GMM (Gaussian Mixture Model), and then calculate the weight and Gaussian distribution parameters of each subspace. The transformation is trained from stereo data of simultaneous recordings of clean and noisy speech, between which, the cepstral degradation is embedded in the statistical relationship. Previous research has shown that SPLICE has a positive effect in improving the recognition rate for ASR and even has a small running cost.

Since the nonlinear transformation model between the clean and noise speech is learned from the training data, however, SPLICE is not effective when the characteristics of the distortion are not known in advance. The inadequately estimated transformations can degrade ASR accuracy.

2.1. Cepstral Enhancement

Through weighted summation of piecewise linear transformations, SPLICE approximates the transformation from $\boldsymbol{y}$ to $\boldsymbol{x}$. Here $\boldsymbol{y}$ is a distorted feature vector and its corresponding clean feature is $\boldsymbol{x}$. We obtain an estimate $\hat{\boldsymbol{x}}$ for $\boldsymbol{x}$ using the method from [8].

$$\hat{\boldsymbol{x}} = \sum_k P(k|\boldsymbol{y})\boldsymbol{A}_k\boldsymbol{y'}, \tag{1}$$

where $\boldsymbol{y'} = \begin{bmatrix} 1 & \boldsymbol{y}^T \end{bmatrix}^T$ and $\boldsymbol{A}_k$ is the linear transformation matrix for subspace k and $\boldsymbol{y'}$ is an augmented feature vector given by $\begin{bmatrix} 1 & \boldsymbol{y}^T \end{bmatrix}^T$. $\boldsymbol{A}_k$ is trained in advance by using stereo data and $P(k|\boldsymbol{y})$ is calculated by using a GMM of distorted features. k is the index of the GMM component.

2.2. SPLICE Training

In the training step for SPLICE, we first learn the probability of distorted features $\boldsymbol{y}$ using GMM as follows,

$$P(\boldsymbol{y}) = \sum_k P(k)P(\boldsymbol{y}|k) = \sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the normal distribution of distorted features $\boldsymbol{y}$. π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ are the weight, the mean and the variance of the k-th component. By obtaining $P(\boldsymbol{y})$ we can determine posterior probability of $P(k|\boldsymbol{y})$ as follows,

$$P(k|\boldsymbol{y}) = \frac{P(k)P(\boldsymbol{y}|k)}{P(\boldsymbol{y})} \tag{3}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \tag{4}$$

The linear transformation matrix $\boldsymbol{A}_k$ is estimated based on the weighted minimum mean square error criterion.

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_i P(k|\boldsymbol{y}_i)||\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}'_i||^2. \tag{5}$$

This estimation needs stereo data, namely, noisy features $\boldsymbol{y}_i$ and their corresponding clean features $\boldsymbol{x}_i$. Because the transformation matrix $\boldsymbol{A}_k$ and the GMM representing noisy features $\boldsymbol{y}$ are trained in advance, the enhancement procedure for SPLICE requires a low computational running cost while achieving high performance. Since $\boldsymbol{A}_k$ is trained using stereo data for available types of noise in the training dataset only, however, the performance of SPLICE will be degraded when input data are given in an unknown noisy environment.

3. EVALUATION USING GOP

GOP is an HMM-based method used to estimate pronunciation proficiency. Previous studies have shown good results using GOP to assess both reading and shadowing speech [6]. GOP provides a score for each phoneme in an utterance. In computing this score, GOP calls for the transcript of the speech to calculate the likelihood. GOP is defined as the posterior probability, $P(p|O^{(p)})$, that the speaker uttered phone p given the corresponding acoustic segment $O^{(p)}$ [4],

$$GOP(p) = \frac{1}{D_p}\log(P(p|O^{(p)})) \tag{6}$$

$$= \frac{1}{D_p}\log(\frac{P(O^{(p)}|p)P(p)}{\sum_{q\in Q} P(O^{(p)}|q)P(q)}) \tag{7}$$

$$\cong \frac{1}{D_p}\log(\frac{P(O^{(p)}|p)}{\max_{q\in Q} P(O^{(p)}|q)}), \tag{8}$$

where, Q is the full set of phonemes, and D_p is the duration of the acoustic segment $O^{(p)}$. It is assumed that the probability of all phonemes is the same, then $P(p) = P(q)$. The sum in the denominator can be approximated as its maximum, and

Table 1. Acoustic model conditions in the ERJ experiment

sampling	16bit/16kHZ
window	Hamming/25 ms
training data	All native speech in the ERJ (5054 utteraces)
parameters	MFCC with CMN, log-energy, and their Δ, $\Delta\Delta$

we can obtain Eq.(8). The numerator in Eq.(8) can be computed via a forced alignment where the sequence of phoneme models is fixed using a given transcript. The denominator can be obtained via continuous phoneme recognition with no constraint.

4. EXPERIMENTS

4.1. Experiment using the ERJ database

The ERJ database contains the clean speech of 190 Japanese learners (college students) and 20 native speakers reading English texts. The English text is divided into 8 sets, each of which contains 60 sentences from TIMIT. Each of the 190 Japanese learners read one set, and each of the 20 native speakers read 4 sets. Note that 2 of the 20 native speakers read all 8 sets. Five of the 60 utterances for each learner have human evaluation scores on 5 point scale, which were given based on how accurately the intended phonemes are acoustically realized. A score was given to each of the five utterances. The human scores were obtained from 5 English native experts in language education, who are also familiar with English spoken by Japanese. The correlation between any two of the 5 English experts' evaluation on average is 0.60.

We trained acoustic models of HMMs (monophones) for GOP using all of the native speech in the ERJ. The acoustic analysis conditions for the GOP scores are shown in Table 1. The testing data used in our experiment include all of the Japanese learners' speech with human scores, for a total of $190 \times 5 = 950$ speech samples and the content of each utterance is different.

Since all the speech samples in ERJ are clean ones, we had to simulate learners' utterances in a noisy condition that are supposed to be observed in a noisy classroom. For this aim, we asked 12 students to read English texts randomly in a classroom and recorded their utterances, which will be used as noise in the experiments. The length of the noise was about 6 minutes. In the recording, the microphone was surrounded by 12 students and the distance from the microphone to each of the students was varied from 2 to 4 meters. Noise caused by adjusting microphone is included in the recording. After recording, the noise file was divided into two 3 minutes long parts. In training SPLICE, the clean speech for SPLICE was the same as the training data used for training native HMMs. To make the stereo-data for SPLICE, a piece of the first noise part was chosen randomly and added to clean data at Signal-to-Noise Ratio (SNR)-5 [dB], SNR0, SNR5, SNR10, SNR15 and SNR20. We used both the simulated noisy speech and the clean speech to perform GMM training using 1024 mixtures.

Table 2. Utterance-level correlations between GOP scores and human scores using the ERJ

SNR	Without SPLICE	With SPLICE
clean	0.550	0.534
SNR20	0.519	0.533
SNR15	0.484	0.517
SNR10	0.417	0.489
SNR5	0.306	0.364
SNR0	0.160	0.195
SNR-5	0.050	0.036

The testing data were also simulated noisy utterances, where noise segments extracted randomly from the second part of the noise file were added to clean speech samples for testing. The correlation of the GOP scores and the teachers' scores, both of which were obtained from testing samples, is compared between the two cases of with and without SPLICE. The experiment was conducted at utterance-level. The results are shown in Table 2.

We can see from Table 2 that the correlations for all noisy speech increased. From SNR20 to SNR5, the lower the SNR is, the larger the correlation improvement is. At SNR10 the increase of correlation is 0.071 and the relative improvement is 17.3%. At SNR5 the increase of correlation is 0.058 and the relative improvement is 21.9%. Since SPLICE may cause a mismatch when it is used for clean data, the correlation for clean data after SPLICE decreases. Similar phenomena are often seen in the case of ASR.

4.2. Experiment using more realistic data

The following three shadowing practice datasets, which are more realistic data than the one used in the previous section, were also used in this research. Note that the ERJ database uses read-aloud utterances, instead of shadowing utterances of this part. Dataset 1 was obtained from a previous paper [6] and was recorded in a quiet classroom. Dataset 2 uses the recordings taken during a college English class in a real environment of a CALL classroom, and Dataset 3 was recorded newly for this paper. The speech materials, which were presented to learners for shadowing practices, were the same among the three datasets. The English proficiency of the learners in three datasets are equally distributed from beginning level to advanced level.

In Dataset 1, 10 utterances of 11 Japanese were recorded in quiet classrooms, but stationary noise that was produced by an air conditioner was added to the utterances. This is a difference in the recording condition from the ERJ database.

Dataset 2 consists of 10 utterances of 12 Japanese college students, which were recorded during English classes in

Table 3. Acoustic modeling conditions for realistic data experiment

sampling	16bit/16kHZ
window	Hamming/25 ms
training data	WSJ+TIMIT HMMs
parameters	MFCC with CMN,log-energy,and their Δ, ΔΔ

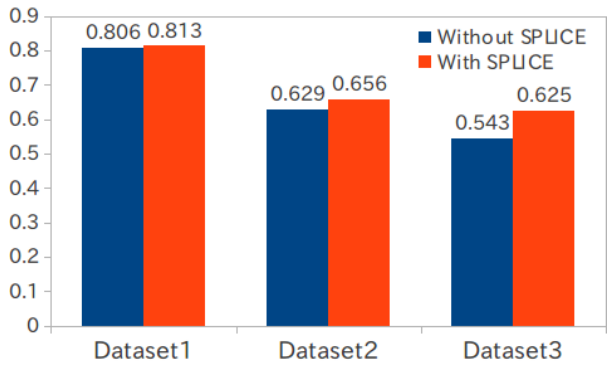

Fig. 1. Utterance-level correlations between GOP score and human score using dataset1, 2,and 3

a CALL room while many other students do shadowing practice at the same time.

Further, we collected a noisier dataset (Dataset 3) of shadowing practice. Dataset 3 contains 9 utterances shadowed by 12 speakers. The 12 speakers in Dataset 3 include two native English speakers, two native Chinese learners, and 8 native Japanese learners. The recording done by letting one person shadow and the others speak English loudly and randomly to make some noise, so the recording environment is much noisier than Dataset 2 and Dataset 1.

The manual assessment of the three datasets was conducted by the same expert (the third author of this paper) in language education. The assessment was done in word unit [6]. If a word in the presented utterance was correctly shadowed, its score is 1. If it is partially correctly shadowed, the score is 0.5. Using this method, every word in the presented utterance comes to have its own score. Furthermore, if unexpected words caused by insertion errors are found in the shadowed utterance, each of the inserted words gives a penalty of -1. The score of a presented utterance is computed by summing up all the scores including the penalty scores. The final score for that utterance is calculated by normalizing the obtained score by the number of the words in the presented utterance.

In this experiment, we used the open source TIMIT+WSJ acoustic models available from the Internet to perform the GOP assessment under the condition of acoustic analysis, shown in Table 3. The SPLICE model used in this experiment is the same as in the experiment using the ERJ database.

The experimental results are shown in Figure 1. We can see from the results that the use of SPLICE improved the correlation between human and machine scores. Dataset 1 was recorded in a quiet classroom, but the relative improvement in correlation is still observed by 0.09%. This means that SPLICE also removes stationary noise such as that generated by an air conditioner. The more noise included in the recordings, the more effective SPLICE comes. In Dataset 2, the relative improvement in correlation is 4.3%. Dataset 3 showed the largest relative improvement, which is 15.3%.

5. SUMMARY AND DISCUSSION

This paper investigated the effectiveness of SPLICE to GOP evaluation performance in a noisy classroom environment. Experimental results show that SPLICE is a highly effective means to improve the GOP performance in a CALL classroom. This paper used both real data and simulated data to perform the evaluations and the improvement is clear for both of them. The noisier the data, the more improvement SPLICE achieves for the GOP evaluation performance.

In the future we plan to evaluate other noise reduction methods such as VTS in the task of proficiency assessment and to test SPLICE and VTS in other tasks such pronunciation error detection in a CALL system.

6. REFERENCES

[1] M.Eskenazi, “An overview of spoken language technology for education,” Speech Communication, 51, 832-844, 2009

[2] L.Chen, “Audio quality issue for automatic speech assessment,” Proc.SLaTE, CD-ROM, 2009.

[3] J.Droppo, L.Deng, and A.Acero, “Evaluation of SPLICE on the Aurora 2 and 3 tasks,”in Proc.ICSLP, 29-32, 2002

[4] S.M. Witt et al., “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech Communication,vol.30, 95-118, 2000.

[5] Sandra Kanters, Catia Cucchiarini, Helmer Strik, “The Goodness of Pronunciation Algorithm: a Detailed Performance Study,” Proc.SLaTE, CD-ROM, 2009

[6] D. Luo, N. Shimomura, N. Minematsu, Y. Yamauchi, and K. Hirose, “Automatic pronunciation evaluation of language learners’ utterances generated through shadowing,” Proc. INTERSPEECH, 2807-2810, 2008

[7] N. Minematsu et al., “Development of English speech database read by Japanese to support CALL research,” Proc. Int. Conf. Acoustics, 557560, 2004

[8] J. Droppo, M. Mahajan, A. Gunawardana, and A. Acero, “How to train a discriminative front end with stochastic gradient descent and maximum mutual information,” Proc. ASRU-2005, 41-46, 2005.