GOP performance improvement of automatic pronunciation assessment in a noisyenvironment *

Yi Luan , Masayuki Suzuki (The University of Tokyo), Yutaka Yamauchi (Tokyo International University), Nobuaki Minematsu, Shuhei Kato , Keikichi Hirose (The University of Tokyo)

1 Introduction

Compared to traditional language education methodologies, CALL systems have many potential benefits. 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 [1]. 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 other learners. Recently, more and more educational facilities have begun to utilize CALL systems during classes. When ASR-based technologies are used, noise from other learners may 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 investigated the effect of using a noise reduction technique in automatic pronunciation proficiency estimation. Here, we tested SPLICE [2], which is a noise reduction algorithm in the presence of additive noise, channel distortion, or a combination of the two. It 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 [3] 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. As well as reading speech, GOP was also used to evaluate the utterances recorded in shadowing practices [5]. We conduct two sets of experiments to evaluate the method. The first set utilized the English Read by Japanese (ERJ) [4] 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. We calculated the correlation between the human scores and the GOP score 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 an average correlation increase of 0.042. The experiment based on real data shows an average correlation increase of 0.041.

The remainder of 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 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, SPLICE separates the space of noisy features into several isolated subspaces according to a GMM (Gaussian Mixture Model), and calculates the weight and normal distribution parameters of each space. The tranformation probability 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 train-

*

ing data, however, SPLICE is not effective when the characteristics of the distortion are not known in advance. The wrong transformation approximation can cause the ASR accuracy to degrade. The results of our experiments show that the noise mismatch between SPLICE training and testing also has a bad influence in GOP-based assessment.

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 [6].

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

Where, $\boldsymbol{y'} = \begin{bmatrix} 1 & \boldsymbol{y}^T \end{bmatrix}^T$ $\boldsymbol{A}_k$ 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 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,

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

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

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 cost while achieving high performance. Since $\boldsymbol{A}_k$ is trained using stereo data for specific type 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. Past studies have shown good results using GOP to assess both reading and shadowing speech [5]. GOP provides a score for each phoneme in an utterance. In computing this score, GOP calls for the transcription 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 $O^{(p)}$ [3],

$$GOP(p) = \frac{1}{D_p} \log(P(p|O^{(p)})) \tag{5}$$

$$= \frac{1}{D_p} \log(\frac{P(O^{(p)}|p)P(p)}{\sum_{q \in Q} P(O^{(p)}|q)P(q)}) \tag{6}$$

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

where, Q is the full set of phonemes, and D_p is the duration of the acoustic segment $O^{(p)}$. Assuming 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 we can obtain Eq.(7). The numerator in Eq.(7) can be computed via a forced alignment where the sequence of phoneme models is fixed using a given transcription. The denominator can be obtained using an unconstrained phoneme loop grammar.

4 Experiments

4.1 Experiment using the ERJ database

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. The human scores were obtained from 5 English native experts in language education, who are also familiar with English spoken by Japanese. The evaluation standard is based on the proficiency of pronunciation, on a 5 point scale. The correlation between the 5 English experts' evaluation on average is 0.57.

We trained acoustic models of HMMs 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 includes 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 (sentence-open).

Since all the speech samples in ERJ are clean ones, we had to simulate learners' utterances in a noisy condition. 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 the 12 students and the distance from

Table 1 Acoustic model conditions in 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, ΔΔ

Table 2 Utterance-level correlations between GOP score and human score 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 microphone to each of the students was varied from 2 [m] to 4 [m]. Noise of 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 synthesized noisy speech and the clean speech to perform GMM training using 1024 mixtures.

The testing data was 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 SNRs were varied from -5 to 20 [dB]. 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 0.07. Since SPLICE may cause a mismatch when it is used for clean data, the correlation for clean data after SPLICE decreases. Similar phenomenon are often seen in the case of ASR.

4.2 Experiment using more realistic data

Three real datasets of shadowing practice were also tested in this research. Dataset 1 was recorded in a quiet classroom. Dataset 2 uses the recordings taken in a college English class, and Dataset 3 was recorded especially for this paper. The speech materials, which were presented to learners for shadowing practices, were the same among the three dataset (sentence-close).

Table 3 Acoustic modeling conditions for real data experiment

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

In Dataset 1, 10 utterances of 11 Japanese ($10 \times 11 = 110$ utterances in total) were recorded in quiet classrooms, but stable noise that was produced by an air conditioner were added to the utterances. This is a difference in recording condition from ERJ database. The TOEIC score for the 12 speakers ranged from 202 to 968 (the full point is 990).

Dataset 2 consists of 10 utterances of 12 Japanese ($10 \times 12 = 120$ utterances in total) college students during English classes in a CALL room while many students do shadowing practice at the same time. The TOEIC score for the 12 speakers ranged from 382 to 870.

In order to prove the effectiveness of SPLICE, we collected a comparatively noisy dataset (Dataset 3) using a shadowing exercise especially for this paper in a classroom. Dataset 3 contain 9 utterances shadowed by 12 speakers ($9 \times 12 = 108$ utterances in total). The 12 speakers in Dataset 3 include two native English speakers, two native Chinese learners, and 8 native Japanese learners. The recording done by having one person shadow and the others were performed to speak English loudly and randomly to make some noise, so the recording environment is much noisier than Dataset 2 or Dataset 1.

The manual assessment was conducted by an expert (the third author of this paper) in language education. The standard of human assessment was done at the word unit [5]. 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 such caused by insertion errors are found in the shadowed utterance, each of the new 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 [8] 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 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

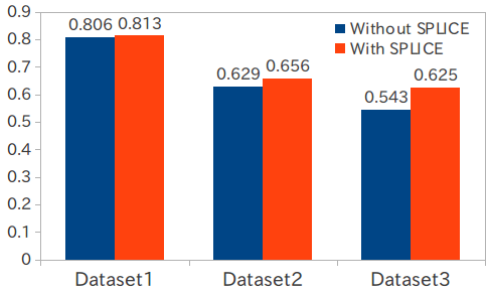

Fig. 1 Utterance-level correlations between GOP score and human score using dataset1, 2,and 3

still improved by 0.007. This means that SPLICE also removes stable noise such as that generated by an air conditioner. The more noise included in the recordings, the more effective SPLICE is. In Dataset 2, the relative improvement in correlation is 0.027. Dataset 3 showed the largest relative improvement at 0.083.

5 Summary and discussion

This paper investigated the application of SPLICE to GOP evaluation performance in a noisy classroom environment. Experimental results show that this is highly effective means to improve GOP performance in this type of situation. SPLICE models the transformation of noisy data to its corresponding clean data as a mixture of Gaussian components for each separate linear space. The merit of SPLICE is that it can model a known type of noise, and this research showed it be very effective in reducing noise caused by other students in classroom.

This paper used both real data and synthesized 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 future we plan to evaluate other noise reduction methods to automatic evaluations and to evaluate SPLICE for use in automatic error detection in a CALL system.

References

[1] L.Chen, "Audio quality issue for automatic speech assessment," in *Proc.SLaTE*, 2009.

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

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

[4] English Speech Database Read by Japanese Students(ERJ database)

[5] 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

[6] 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.

[7] Y. Qiao and N. Minematsu, "Mixture of probabilistic linear regressions: a unified view of GMM-based mapping techniques," Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP 2009), 3913-3916, 2009

[8] Keith Vertanen, "WSJ+TIMIT recipe," http://www.keithv.com/software/htk/