AUTOMATIC INTELLIGIBILITY ASSESSMENT AND DIAGNOSIS OF CRITICAL PRONUNCIATION ERRORS FOR COMPUTER-ASSISTED PRONUNCIATION LEARNING

Antoine RAUX and Tatsuya KAWAHARA

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{raux, kawahara}@kuis.kyoto-u.ac.jp

ABSTRACT

We introduce a novel method to diagnose pronunciation errors that are most critical to the intelligibility of L2 learners. A preliminary study showed that error rates computed by a speech recognition-based system can be used to characterize intelligibility. We deduce a probabilistic algorithm to derive intelligibility from error rates. We also define an error priority function that indicates which errors are most critical to intelligibility. Experimental results proved the validity of the approach.

1. INTRODUCTION

Recent Computer-Assisted Pronunciation Learning (CAPL) research, stirred up by the improvement of computer hardware and Automatic Speech Recognition (ASR), has focused on two areas: evaluation and instruction. While studies on the correlation between acoustic features of speech and human judgements of intelligibility opened the way to automatic evaluation of intelligibility ([1],[2]), innovative approaches to instruction were developed using ASR to detect segmental and prosodic errors ([3],[4]). However, little has been done to relate instruction and evaluation, and to provide learners with feedback on which errors are most critical to their intelligibility. This may result in sub-optimal learning as students spend time on aspects of pronunciation that do not noticeably affect intelligibility. In this paper, we propose a new method to assess intelligibility from pronunciation error rates and spot the errors that are most critical to each learner's intelligibility.

2. RELATIONSHIP BETWEEN ERROR RATES AND INTELLIGIBILITY

2.1. Experimental Protocol

We conducted an experimental study of the relationship between 10 selected pronunciation errors common among Japanese speakers of English and human ratings of intelligibility. The list of errors is given in Table 1. There were 16 subjects, all of whom were students, faculty or staff members from Kyoto University. We recorded their reading of a passage designed for pronunciation evaluation[5]. The recordings were sent to a qualified linguist who rated each subject's intelligibility from 1 (hardly intelligible) to 5 (perfectly intelligible). Pronunciation errors in the recordings were then detected using ASR, and each subject's error rates were computed. We computed the average error rates of subjects of each intelligibility level. Figure 1 shows the error rates of the 5 levels.

2.2. Results

Figure 1 shows that the way error rates vary accross levels depends on the error. Students of different levels are grouped according to their performance on different error categories. Three types of errors can be distinguished as follows.

2.2.1. Phonemic substitutions and deletion

For these errors (number 1 to 4), only level-5 students have a low error rate[1] . The other students have an error rate of about 10 to 30% greater depending on the error. The error rates among students of level 1 to 4 are equivalent.

2.2.2. Vowel non-reduction

This error (number 5) divides the students into 3 groups: level-1 students have an error rate of 90%, students of level 2 to 4 have an error rate of $60 - 70\%$ and students of level 5 students have an error rate of 40%.

2.2.3. Vowel insertion, H/F, V/B substitutions

These errors (number 6 to 10) include syllable structure and two consonant contrasts (/v-b/ and /h-f/). Level-1 students

[1]For all errors except error 6 ($p < 0.2$), all differences between average error rates are significant at the 0.05 level.

Table 1. Errors detected by the system

Number	Description	Example word	Erroneous pronunciation
1	Word-initial w/y deletion	would	uː d
2	SH/CH substitution	choose	ʃ uː z
3	ER/A substitution	paper	p eI p ɑː
4	R/L substitution	road	l o ʊ d
5	Vowel non-reduction	student	s t j uː d e n t
6	V/B substitution	problem	p r ɑ v l ə m
7	Word-final vowel insertion	let	l e t ɔː
8	CCV-cluster vowel insertion	study	s u t ʌ d i
9	VCC-cluster vowel insertion	active	ɑ k u t i v
10	H/F substitution	fire	h aI $ə^r$

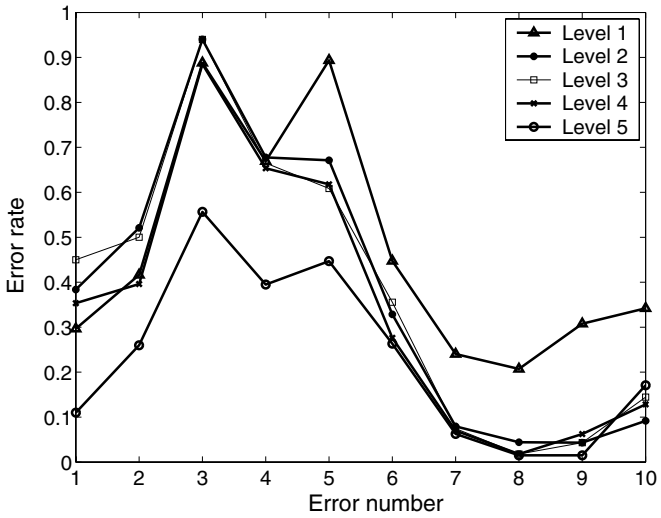

Fig. 1. Error rates averaged for each intelligibility level.

have larger error rates than other students. Students of level 2 to 5 have similar low error rates.

2.3. Linguistic Interpretation

These observations show that such aspects of pronunciation as consonant clusters (affected by vowel insertion) or vowel reduction need to be mastered in order to reach even average levels of intelligibility. On the contrary, phonemic substitutions and deletions, except for the two pairs H/F and V/B, do not prevent speakers to be largely intelligible, since even largely intelligible speakers (level 4) have high error rates. These results are consistent with the position of most recent linguists regarding the teaching of pronunciation[6]. Errors such as vowel insertion and non-reduction which are related to prosodic features such as syllable structure and stress are considered to be more crucial to intelligibility than purely segmental errors.

3. INTELLIGIBILITY ASSESSMENT

3.1. Probabilistic Models

Based on the findings of the preliminary study, we propose a probabilistic approach to intelligibility assessment. Given observed error rates O, our goal is to obtain the probability that the learner's intelligibility level is i, ($i \in \{1..5\}$). This probability, noted $P(i|O)$, can be computed using Bayes formula:

$$P(i|O) \propto P(i)P(O|i) \tag{1}$$

where $P(i)$ is the ratio of level-i students in the considered population and $P(O|i)$ is the probability distributions of the error rates for level-i speakers. Under the assumption that all error rates are statistically independent *given the intelligibility level*, the overall probability distribution is given by $P(O|i) = \prod_j P(r_j|i)$, where $P(r_j|i)$ is the probability distribution of the jth error rate among students of level i. We model each $P(r_j|i)$ by a Beta distribution, defined on $[0, 1]$ by:

$$\beta_{(a,b)}(x) = B(a, b)x^{(a-1)}(1 - x)^{(b-1)} \tag{2}$$

where a and b are parameters and $B(a, b)$ is a normalizing constant. Parameters are computed using data rated for intelligibility by a human judge. Combining equations 1 and 2 leads to the following formula for the probability of level i:

$$P(i|O) = K \prod_j \beta_{(a,b)_{i,j}}(r_j) \tag{3}$$

where K is a normalizing constant. We define the intelligibility score as the expected value of the level:

$$I = \sum_i i \cdot P(i|S) \tag{4}$$

Thus, the score can take any value in the range $[1, 5]$.

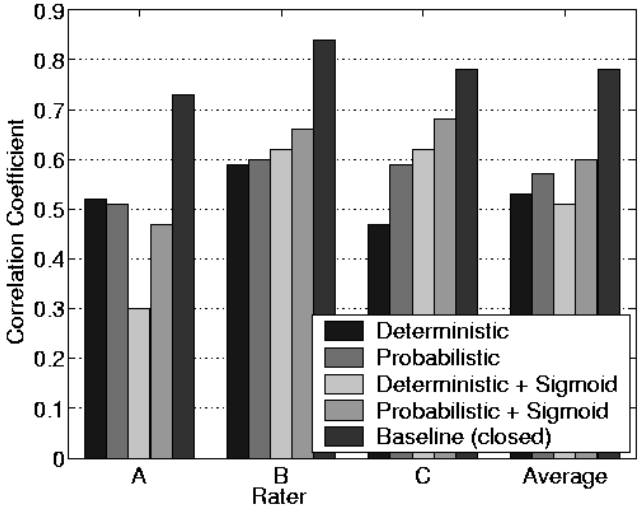

Fig. 2. Correlation coefficients on cross-validation

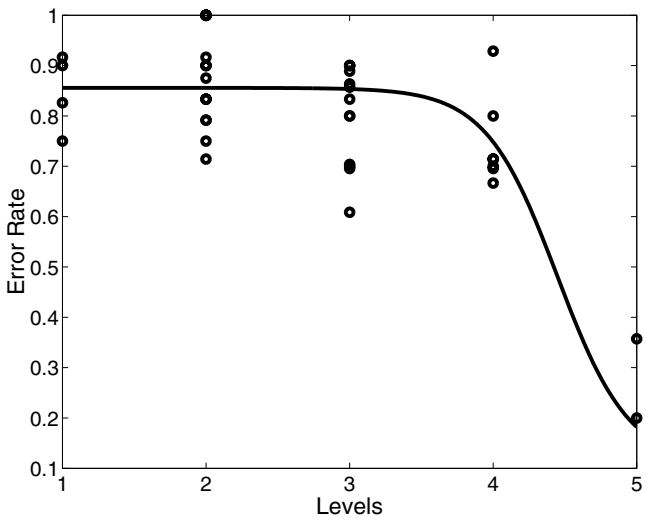

Fig. 3. Data points and fitted sigmoid for ER/A error

3.2. Evaluation

3.2.1. Speech data and intelligibility ratings

We evaluated the algorithm on data containing both reading passages and prepared oral presentations, featuring a total of 42 Japanese subjects. Each subject was rated for intelligibility by 3 associate professors of English at Japanese universities. Inter-rater correlations were surprisingly low, even after normalizing the scores to compensate for individual raters' bias (0.6 on average). Therefore, evaluation was conducted on each rater separately.

3.2.2. Closed evaluation

We trained the model (i.e. computed the error rate distributions for each level) on the whole set of speakers and estimated the level of each speaker. The correlations between the each rater and the corresponding machine score were respectively 0.73, 0.84 and 0.78 (the rightmost bars in Figure 2), confirming the validity of the approach.

3.2.3. Take-one-out evaluation

In this case, we trained the model using 41 subjects and estimated the intelligiblity of the remaining subject, and repeated this 42 times. Correlations were much lower, respectively 0.52, 0.59 and 0.47. As well as the high variability of the evaluation data, the fact that some levels contained very few subjects (< 4) makes it harder to reliably estimate probability distributions and degrades the performance of the system.

3.3. Smoothing the Distributions

3.3.1. Modelling the uncertainty of the error rates

To compensate for these lack of data, we tried two approaches: using a probabilistic model of error rates and constraining the error rates' probability distributions. In the former case, instead of considering the error rate as a deterministic variable, we model it by a belief distribution on $[0, 1]$ whose mean is the measured error rate and whose variance depends on the number of observed patterns (the more observations, the smaller the variance). We used Beta density functions to model them as well. Calculus yields an exact formula for $P(i|O)$ as well, which replaces equation (3):

$$P(i|O) \quad = \quad KP(i) \prod_j \frac{B(a_j + a'_j, b_j + b'_j)}{B(a_j, b_j)} \qquad (5)$$

where a'_j and b'_j are parameters of the belief distribution.

3.3.2. Constraining the distributions

The second improvement is based on the assumption that all error rates must decrease as intelligibility increases. We apply this by constraining the rates of each error to be on a decreasing sigmoid function of the level, instead of computing the exact average from the training data. The sigmoid is fitted on the data by gradient descent using mean squares error. Figure 3 gives an example of a sigmoid function fitting the training data points for ER/A substitution (error 3).

The results of these modifications are summarized in Figure 2. Both modifications improved the correlations, the best result being an average correlation of 0.6, obtained by combining the two.

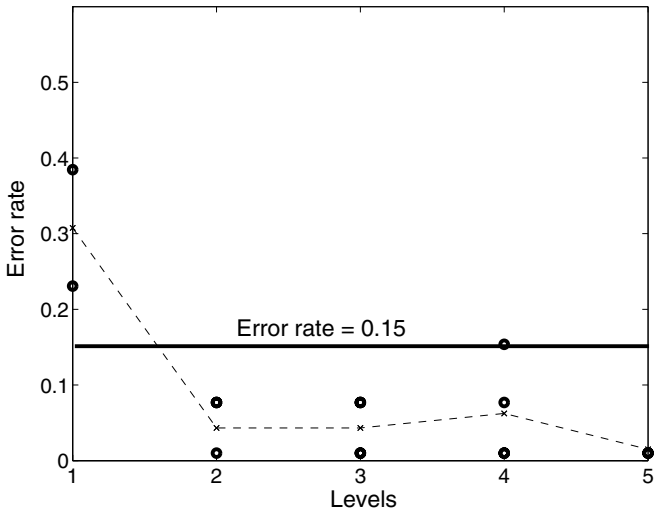

Fig. 4. Average error rates and priority for VCC vowel insertion (because some points are superimposed, less than 16 points appear).

4. DIAGNOSIS OF CRITICAL PRONUNCIATION ERRORS

4.1. Error Priority

To determine which errors should be studied by a given learner, we define the priority $\pi(j,i)$ of error j at the intelligibility level i as the difference between the learner's error rate and the average error rate of level-i students, that is:

$$\pi(j,i) = r_j - \langle r_j \rangle_{level-i\ students} \quad (6)$$

The priority $\pi(j)$ of error j is defined as the expected value of each level's priority:

$$\pi(j) = \sum_i P(i|O) \cdot \pi(j,i) \quad (7)$$

4.2. Example

Figure 4 gives an illustration of how error priority is affected by the overall intelligibility of the student. In this example, we consider a student whose error rate on error 9 (vowel insertion in VCC clusters) is 0.15. Circles represent students from the training samples and the dashed line connects the average value for each intelligibility level. The distance between the horizontal line and the dashed line represents the absolute value of the priority for this level. Priority is negative for level 1 (the student's rate is below the average) and positive for all other levels. Thus, this error is likely to be proposed for study if the learner's intelligibility level is 2 or more, because speakers of these levels usually master this error. On the other hand, if the learner's level is 1, other errors are more likely to be proposed to improve intelligibility.

We computed error priorities for the 16 speakers of the preliminary experiment and found that they agreed with subjective judgements of the strengths and weaknesses of each speaker. To conduct a formal evaluation, we have developed a prototype CAPL system that uses the algorithms presented in this paper. Using this prototype, we hope to demonstrate the validity of this algorithm and its applicability to real world situations.

5. CONCLUSION

We proposed a probabilistic method to assess non-native speakers' intelligibility. Our approach yields an explicit model of the relationship between intelligibility and error rates. We showed that it can be used to provide meaningful feedback to the learners on their strengths and weaknesses. Although performance needs to be improved to be applied to practical systems, promising results were obtained, which could find applications in CAPL systems, and more generally, Intelligent Tutoring Systems.

Acknowledgments

We would like to thank Prof. Hiroshi Okuno, Prof. Masatake Dantsuji, Dr. Rebecca Dauer and Yasushi Tsubota for their help and advice on this research.

6. REFERENCES

[1] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *ICASSP'97*, 1997, vol. 2, pp. 1471–1474.

[2] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of dutch pronunciation by using speech recognition technology," in *IEEE workshop ASRU*, 1997, pp. 622–629.

[3] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji, "Japanese pronunciation instruction system using speech recognition methods," *IEICE Trans.*, vol. E83-D, no. 11, pp. 1960–1968, 2000.

[4] M. Eskenazi, "Using computer in foreign language pronunciation training: What advantages?," *CALICO Journal*, vol. 16, no. 3, 1999.

[5] C.H. Prator and B.W. Robinett, *Manual of American English Pronunciation*, HRW International Editions, 1985.

[6] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin, *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages*, CUP, 1996.