New feature parameters for pronunciation evaluation in English presentations at international conferences

Hiroshi Kibishi and Seiichi Nakagawa

Department of Computer Science and Engineering
Toyohashi University of Technology, Japan
{kibishi, nagkagawa}@slp.cs.tut.ac.jp

Abstract

We have previously proposed a statistical method for estimating the pronunciation proficiency and intelligibility of presentations made in English by non-native speakers. To investigate the relationship between various acoustic measures and the pronunciation score and intelligibility, we statistically analyzed the speaker's actual utterances to find combinations of acoustic features with a high correlation between the score estimated by a linear regression model and the score perceived by native English teachers. In this paper, we examined the quality of new acoustic features that are useful when used in combination with the system's estimates of pronunciation score and intelligibility. Results showed that the best combination of acoustic features produced correlation coefficients of 0.929 and 0.753 for pronunciation and intelligibility, respectively, using open data for speakers at the 10-sentence level.

Index Terms: pronunciation evaluation, intelligibility evaluation, English, HMM, phoneme pair, perplexity

1. Introduction

Many researchers have investigated automatic methods for evaluating pronunciation proficiency. Nuemeyer et al., for example, proposed an automatic text-independent pronunciation scoring method. They used an HMM log-likelihood score, segment classification error scores, segment duration scores, and syllabic timing scores for the French language [1]. They found that evaluation by segment duration showed better results than other methods. Franco et al. investigated an evaluation method based on an HMM-based phone log-posterior probability score and a combination of the scores [2]. We also previously investigated the use of the posterior probability as an evaluation measure [3]. Furthermore, Franco et al. proposed using the log-likelihood ratio score of native acoustic models to non-native acoustic models and found that this measure outperformed posterior probability evaluation [4].

Cucchiarini et al. compared the acoustic scores as measured by TD (total duration of speech including pauses), ROS (rate of speech; total number of segments/TD), and LR (a likelihood ratio, corresponding to the posterior probability) and showed that TD and ROS correlated more highly with human ratings than LR [5].

These above mentioned studies all focused on either European languages or English uttered by European non-native speakers. In contrast, we have evaluated Japanese uttered by foreign students in Japan [8].

In our earlier work we proposed a statistical method for evaluating the pronunciation proficiency of Japanese speakers giving presentations in English [7][9][10].

In this paper, we build on this previous research by proposing a statistical method for estimating the pronunciation score and intelligibility of presentations given in English by Japanese speakers. Because automatic transcription rates in phoneme and word recognition are not directly related to intelligibility, we investigated the relationship between pronunciation score / intelligibility and various acoustic measures, and then combined these measures using a linear regression model. Finally, we examined the effectiveness of new acoustic features such as perplexity and phoneme pair discrimination rate when the system estimated the pronunciation score and intelligibility. As far as we know, the automatic estimation of intelligibility has not yet been studied.

2. Database and system overview

We used the Translanguage English Database (TED)[16], presented at EuroSpeech, for evaluating the test data. Only part of the TED has transcribed texts. This data set consists of $21(speakers) \times 10 \sim 21(sentences)$, giving a total of 289 English sentences appearing in speeches given by 21 male speakers with above average, average, or below average pronunciation proficiency. 16 of the 21 were native Japanese speakers, while the other five were native English speakers from the United States. We used the TIMIT/WSJ database for training native English phoneme HMMs, a separate Japanese speech database for adapting them (non-native English phoneme HMMs)[6], and the ASJ/JNAS database for training native Japanese syllable HMMs (strictly speaking, mora-unit HMMs).

Table 1 gives a summary of the speech material. All speech was downsampled to 16kHz and pre-emphasized, after which a Hamming window with a width of 25 ms was applied every 10 ms. Twelve-dimensional MFCCs were used as the speech feature parameters for each frame. The acoustic features were the 12 MFCCs, Δ and $\Delta\Delta$ features. Acoustic models based on monophone syllable HMMs were trained based on the analyzed speech. The English monophone HMMs are composed of three states, each of which has four mixture Gaussian distributions with full covariance matrices. The Japanese syllable HMMs are composed of four states, each of which has four mixture Gaussian distributions with full covariance matrices.

Witt et al. found that for the pronunciation evaluation of non-native English speakers, triphones perform worse than monophones if the HMMs are trained by native speech; in other words, less detailed (native) models perform better for non-native speakers [11][12][13].

Figure 1 shows a block diagram of our evaluation system for pronunciation score and intelligibility. Acoustic feature measures are extracted from presentations given during lectures, and both scores are estimated by their corresponding regression models.

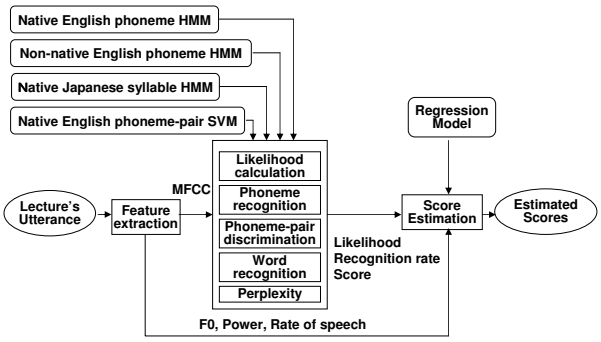


Figure 1: Block diagram of our estimation system for pronunciation score and intelligibility.

Table 1: Speech material for training HMMs.

HMM	speaker (database)	♯speakers	♯total sentences
English	Native (TIMIT)	326	3260
	(WSJ)	50	6178
	Japanese students	76	1065
Japanese	Native (ASJ)	30	4518
	(JNAS)	125	12703

3. Pronunciation score and intelligibility rated by English teachers

3.1. Definition of pronunciation score

The pronunciation score used in this paper is the average of two scores: a phonetic pronunciation score and a prosody (rhythm, accent, intonation) score, rated by five American English teachers for each of the 289 sentences.

3.2. Definition of inte5lligibility

Intelligibility in this paper is defined as how well English teachers recognize the pronunciation of non-native speakers.

Four American English out of the above teachers transcribed each sentence while scoring the speaker's pronunciation proficiency. The 4 transcriptions of the same sentence were compared, and any word correctly transcribed by 2 or more English teachers was referred to as **man2/4**. After computing all the man2/4 values for all utterances, intelligibility could be calculated as:

$$Intelligibility = A/B, \quad (1)$$

where A represents the number of words transcribed as man2/4 in each sentence, that is, how many words a teacher recognized, and B represents the total number of words in each sentence. However, because we did not have transcriptions of the test data from the speakers themselves, we were not able to obtain the exact number of words in the sample sentences. Consequently, we assumed the total number of words in a sentence to be the sum of the number of words transcribed as man2/4 in the sentence combined with the average number of transcribed words not included in the man2/4 figures from the same sentence.

4. Definition of measures for each acoustic feature

In this paper, other than the features that we have used thus far ((a)∼(i))[7][9][10], we add the new features of perplexity, entropy, spectrum rate, and phoneme pair discrimination score.

4.1. Acoustic Features

(a) Log-likelihood using native and non-native English HMMs and the learner's native language HMM
We calculated the correlation rate between the observed scores and the log-likelihood (LL) for a pronunciation dictionary sequence based on the concatenation of phone HMMs at every 1, 5, and 10 sentence levels. The likelihood was normalized by length in the frames. We used both native English phoneme HMMs (LL_{native}) and non-native English phoneme HMMs adapted based on Japanese utterances ($LL_{non-native}$).

(b) Best log-likelihood for arbitrary phoneme sequences
The best log-likelihood for arbitrary phoneme sequences is defined as the likelihood of arbitrary phoneme (syllable) recognition without using phonotactic language models. We used native English phoneme HMMs (LL_{best}) .

(c) Log-likelihood ratio
We used the log-likelihood ratio (LR) between native English HMMs and non-native English HMMs, which is defined as the difference between the two log-likelihoods, that is, $LL_{native} - LL_{non-native}$.

(d) A posteriori probability
We used the likelihood ratio (LR') between the log-likelihood of native English HMMs (LL_{native}) and the best log-likelihood for arbitrary phoneme sequences (LL_{best}), which means the *a posteriori* probability, that is, $LL_{native} - LL_{best}$ [7].

(e) Likelihood ratio for phoneme recognition
We used the ratio of the likelihood of arbitrary phoneme recognition between native English HMMs and non-native English HMMs (LR_{adap}), which is defined as the difference between the two log-likelihoods, that is, $LL_{best\_native} - LL_{best\_non-native}$.

We also used the ratio of the likelihood of arbitrary phoneme (syllable) recognition between native English HMMs and native Japanese syllable HMMs (LR_{mother}), which is defined as the difference between the two log-likelihoods, that is, $LL_{best\_native} - LL_{best\_mother}$.

(f) Phoneme recognition results
We used the correct rate, substitution rate, and deletion rate of arbitrary phoneme recognition. The test data were restricted to the correctly transcribed parts according to the man2/4 transcriptions.

(g) Word recognition result
The correct rate of word recognition was used with a language model. We used the WSJ database (WSJ) and a Eurospeech '93 paper (EURO) for training bigram language models [9]. The test data were limited to the correctly transcribed parts according to the man2/4 transcriptions.

(h) Standard deviation of powers and F_0
We calculated the standard deviation of the powers ($Power$) and fundamental (pitch) frequencies (F_0).

(i) Rate of speech (ROS)
We used the rate of speech of the sentence. Silences in utterances were removed. We calculated the ROS of each sentence as the number of phonemes divided by the duration in seconds.

(j) Perplexity
Perplexity can be used to evaluate the complexity of an utterance. We used the WSJ database (WSJ) and a Eurospeech '93 paper (EURO) for training bigram language models [9]. En-

tropy H and perplexity PP can be calculated for a word sequence $w_1 w_2 \cdots w_n$ in a test set:

$$H = -\frac{1}{n} log_2 p(w_1 \cdots w_n) \quad (2)$$

$$PP = 2^H \quad (3)$$

As for cases of out-of-vocabulary, adjusted perplexity, they can be calculated as:

$$APP = (P(w_1 \cdots w_n) m^{n_\mu})^{-\frac{1}{n}} \quad (4)$$

where n_μ represents the number of out-of-vocabulary, and m represents the number of kinds of out-of-vocabulary items in a test set.

(k) Spectrum changing rate
A native speaker's English utterances are spontaneous, and the spectrum changing rate may therefore vary rapidly. Spectrum changing rate can be calculated as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)} \quad (5)$$

We examine Euclid distance between adjacent frames of calculated MFCC, and we use the standard variation and variance. Where i represents the i-th index, x_i represents MFCC of i dimension, and y_i represents MFCC in the previous frame of the i dimension.

(l) Phoneme pair discrimination score
We discriminated nine pairs of phonemes by SVM that are often mispronounced by Japanese native speakers. They are / l and r /, / m and n /, / s and sh /, / s and th /, / b and v /, / b and d /, / z and dh /, / z and d /, and / dh and d /.

The SVM input data comprised fixed length frames, that is, five consecutive frames beginning from the -2 frame of the central frame of the phoneme segment. The features are MFCC and Δ MFCC.

The phoneme pair discrimination score is a value that reflects a quantized distinction rate from 1 to 4 for every sentence. Each sentence includes an average of 37 phoneme pairs. The average correct discriminative ratios of native English phonemes and Japanese English phonemes were 89.0% and 79.3%, respectively.

4.2. Correlation between the scores and measures of each acoustic feature

Tables 2 and 3 summarize the correlation between each acoustic measure and the English teacher pronunciation scores/intelligibility, respectively. The lower rows (bold) correspond to the newly added features. For the pronunciation score, the correlations of the spectrum changing rate and phoneme pair discrimination score are high. As for perplexity, we expected that a speaker with good pronunciation might utter a complicated sentence and unfamiliar words, for which a positive correlative value would be observed, but the results showed a negative value. This result indicates that pronunciation scores and intelligibility become worse when a speaker utters a complicated sentence and unfamiliar words.

Among the conventional acoustic features, LR was found to have the highest correlation values.

Table 2: Correlation between acoustic measures and pronunciation scores

Measure	1 sentence	5 sentences	10 sentences
LL_{native}	-0.466	-0.601	-0.626
LR	0.800	0.877	0.905
LR'	0.214	0.321	0.382
Phoneme recog ($Cor.$)	0.299	0.461	0.506
Word recog (EURO, $Cor.$)	0.113	0.242	0.289
$Rate\ of\ speech$	0.523	0.700	0.753
$\boldsymbol{PP}$**(EURO)**	-0.077	-0.187	-0.257
$\boldsymbol{PP}$**(WSJ)**	-0.068	-0.151	-0.203
$\boldsymbol{APP}$**(EURO)**	-0.077	-0.187	-0.256
$\boldsymbol{APP}$**(WSJ)**	-0.051	-0.112	-0.145
$\boldsymbol{H}$**(EURO)**	-0.007	-0.029	-0.077
$\boldsymbol{H}$**(WSJ)**	-0.298	-0.574	-0.719
Spectrum changing rate	0.320	0.339	0.329
Spectrum rate (SD)	0.400	0.517	0.578
Spectrum rate (variance)	0.413	0.532	0.592
Phoneme-pair	0.241	0.462	0.590

Table 3: Correlation between acoustic measures and intelligibility

Measure	1 sentence	5 sentences	10 sentences
LL_{native}	-0.180	-0.389	-0.527
LR	0.184	0.421	0.496
LR'	0.337	0.449	0.546
Phoneme recog ($Cor.$)	-0.117	0.083	0.266
Word recog (EURO, $Cor.$)	0.009	0.248	0.220
$Power$	-0.022	-0.131	-0.197
$Pitch(F_0)$	0.196	0.353	0.455
$Rate\ of\ speech$	0.166	0.309	0.354
$\boldsymbol{PP}$**(EURO)**	-0.113	-0.188	-0.121
$\boldsymbol{PP}$**(WSJ)**	0.041	-0.006	0.024
$\boldsymbol{APP}$**(EURO)**	-0.113	-0.188	-0.120
$\boldsymbol{APP}$**(WSJ)**	0.045	0.024	0.085
$\boldsymbol{H}$**(EURO)**	-0.052	-0.080	-0.047
$\boldsymbol{H}$**(WSJ)**	-0.047	-0.234	-0.461
Spectrum changing rate	0.197	0.339	0.404
Spectrum rate (SD)	0.098	0.160	0.245
Spectrum rate (variance)	0.101	0.168	0.255
Phoneme-pair	0.132	0.340	0.503

5. Statistical method for estimating pronunciation score and intelligibility

For estimating the pronunciation score and intelligibility, we propose a linear regression model, derived from the relationship between the observed acoustic measures and the English teacher scores. Having established some independent variables $\{x_i\}$ for the parameters and the value Y for the English teacher scores, we define the linear regression model as

$$Y = \Sigma_i \alpha_i \times x_i + \varepsilon, \quad (6)$$

where ε is the residue [7][8]. The coefficients $\{\alpha_i\}$ are determined by minimizing the square of ε. We conducted experiments with open data for speakers. We investigated whether our proposed method functions independently of the speaker. For an open experiment using these speakers, we estimated a regression model using the utterances of 20 of the speakers and estimated the score of the remaining speaker. We repeated this procedure for every speaker.

Table 4: Correlation between the combination of acoustic measures and pronunciation scores rated by humans
"bold" denotes the new features proposed here

Acoustic measures / Number of sentences for evaluation	1 sentence	5 sentences	10 sentences
Word recog(EURO, $Cor.$), LR, $Power$, Word recog(WSJ, $Cor.$)	0.770	0.866	0.884
Word recog(EURO, $Cor.$), LR, $Power$	0.771	0.858	0.887
LL_{native}, $LL_{non-native}$, LR, LR_{mother}, $Power$, Phoneme recog($Del.$), **H(WSJ), Phoneme−pair**	**0.807**	0.862	0.867
Word recog(EURO, $Cor.$), LR, $Power$, Word recog(WSJ, $Del.$), Word recog(WSJ, $Cor.$), **H(WSJ), APP(EURO), PP(EURO), Phoneme−pair**	0.751	**0.881**	**0.929**

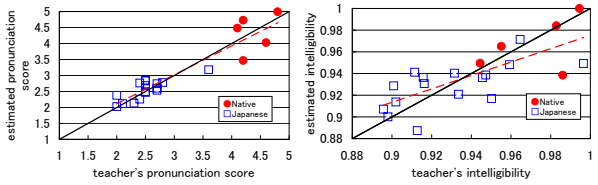
Table 5: Correlation between the combination of acoustic measures and intelligibility rated by humans
"bold" denotes the new features proposed here

Acoustic measures / Number of sentences for evaluation	1 sentence	5 sentences	10 sentences
LR', Word recog(EURO, $Cor.$), LR_{adap}, $Power$, Phoneme recog($Cor.$)	0.347	0.550	0.624
LL_{native}, LR', Phoneme recog($Sub.$), Phoneme recog($Del.$), Phoneme recog($Cor.$), LR	0.225	0.274	0.724
$LL_{non-native}$, LL_{best}, $Pitch(F_0)$, $LRmother$, $LRadap$, Phoneme recog($Cor.$), **APP(WSJ)**	**0.476**	0.518	0.499
LR', Word recog(EURO, $Cor.$), LR_{adap}, $Power$, Phoneme recog($Cor.$), **Spectrum changing rate(average), Phoneme−pair**	0.356	**0.652**	0.752
$LL_{non-native}$, LR', Phoneme recog($Sub.$), **PP(WSJ), PP(EURO), APP(EURO), Phoneme−pair**	0.129	0.537	**0.753**

Tables 4 and 5 summarize the results of the pronunciation scores and intelligibility, respectively, for the open data obtained at levels with 1, 5, and 10 sentences. By combining certain acoustic measures, we obtained correlation coefficients of 0.929 and 0.753 for the pronunciation scores and intelligibility, respectively, using open data with each speaker at a level with 10 sentences.

Figure 2 illustrates the relationship between the estimated pronunciation score and intelligibility and that of the English teachers in based on the open data for a set of 10 sentences.

These results confirm that the proposed method for automatically estimating pronunciation scores and intelligibility has approximately the same effectiveness as actual evaluations performed by English teachers.



(a) pronunciation score (b) intelligibitily

Figure 2: Relationship between estimation scores and teacher scores

6. Conclusion

In this paper, we proposed a statistical method for estimating the pronunciation score and intelligibility of presentations made in English by non-native speakers based on a linear regression model. By combining some new acoustic and linguistic measures, our proposed method was able to evaluate the pronunciation score and intelligibility with almost the same accuracy and effectiveness as actual English teachers.

In our future development of the method, we plan to develop a system which identifies unsuitable phoneme pronunciations for users based on phoneme pair discrimination results.

7. References

[1] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. ICSLP*, pp.1457-1460, 1996.

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

[3] Y. Taniguchi, A.A. Reyes, H. Suzuki, and S. Nakagawa, "An English conversation and pronunciation CAI system using speech recognition technology," in *Proc. EuroSpeech*, pp.705-708, 1997.

[4] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Proc. EuroSpeech*, pp.851-854, 1999.

[5] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," in *Speech Communication*, 30(2-3), pp.109-119, 2000.

[6] S. Nakagawa, Allen A. Reyes, H. Suzuki, and Y. Taniguchi, "An English conversation CAI system using speech recognition technology," Trans. Information Processing Society in Japan. Vol.38 No. 8, pp. 1649-1657 (1997, in Japanese)

[7] S. Nakagawa, N. Nakamura, and K. Mori, "A statistical method of evaluating pronunciation proficiency for English words spoken by Japanese," IEICE Trans. Inf. & Syst., vol.E87-D, no.7, pp1917-1922, July 2004

[8] K. Ohta and S. Nakagawa, "A Statistical Method of Evaluating Pronunciation Proficiency for Japanese Words," in *Proc. Interspeech*, pp.2233-2236, 2005.

[9] K. Ohta and S. Nakagawa, "A Statistical Method of Evaluating Pronunciation Proficiency for Presentation in English," in *Proc. Interspeech*, pp.2317-2320, 2007.

[10] K. Hirabayashi,and S. Nakagawa, "Automatic Evaluation of English Pronunciation by Japanese Speakers Using Various Acoustic Features and Pattern Recognition Techniques," in *Proc. Euro Speech*, pp.598-601, 2010.

[11] He,X., Zhao, Y., "Model complexity optimization for nonnative English speakers.", in *Proc. EuroSpeech*, pp.1461-1463, 2001.

[12] Ronen, O., Neumeyer, L.,Frando, H. "Automatic detection of mispronunciation for Language Instruction.", in *Proc. EuroSpeech*, Vol. 2, pp.649-652, 1997.

[13] Witt, S., Young, S. "Offline acoustic modeling of non-native accents.", in *Proc. EuroSpeech*, Vol. 3, pp.1367-1370, 1999.

[14] Acoustical Society of America "SII: Speech Intelligibility Index", http://www.sii.to/index.html

[15] Advanced Utilization of Multimedia to Promote Higher Education Reform "English Speech Database Read by Japanese Students", http://research.nii.ac.jp/src/eng/list/detail.html

[16] "TED Translanguage English Database", http://www.elda.org/catalogue/en/speech/S0031.html