INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Improvement of Automatic English Pronunciation Assessment with Small Number of Utterances Using Sentence Speakability

Satsuki Naijo, Akinori Ito, Takashi Nose

Graduate School of Engineering, Tohoku University, Japan
{satsuki.naijo.t2@dc., takashi.nose.b7@, akinori.ito.a2@}tohoku.ac.jp

Abstract

The current Computer-Assisted Pronunciation Training (CAPT) system uses DNN-based speech recognition results to evaluate learner's pronunciation with high accuracy when using many utterances for the evaluation. However, when we use only a few utterances, the accuracy of the CAPT system deteriorates. One reason for the deterioration is that the score calculated by a CAPT system is biased depending on the pronunciation difficulty of the sentences when using a small number of utterances. In this study, we developed a CAPT system that takes the sentence speakability (pronunciation difficulty of sentences) into account. As a result, the correlation coefficient between the human evaluation and the machine score was 0.46 in the conventional method, while it improved to 0.57 with the proposed method.

Index Terms: computer-assisted pronunciation training, speech recognition, sentence speakability

1. Introduction

In recent years, the need for students to acquire English language skills has increased due to globalization. Computer-Assisted Language Learning (CALL) systems have been attracting attention as a learning platform. A system for learning English pronunciation is called the Computer-Assisted Pronunciation Training (CAPT) system, which uses speech recognition technology to recognize the learner's speech and evaluate pronunciation based on the recognition results[1].

The CAPT system measures how well the accuracy of pronunciation evaluation correlates with human evaluation. A variety of scores have been devised to improve accuracy. A typical example is the Goodness of Pronunciation (GOP), calculated from the likelihood obtained from the hidden Markov models (HMMs) with forced-alignment calculation [2][3][4]. It evaluates the learner's pronunciation at the phoneme level. Next, a method using speech recognition with hidden Markov Model with Gaussian mixture model (GMM-HMM) [5][6]has been proposed to evaluate the whole utterance and has obtained a high correlation with human evaluation. In recent years, pronunciation evaluation using deep neural network-based speech recognizers (DNN-HMMs) has also emerged and has achieved higher accuracy than previous pronunciation evaluations [7][8][9]. However, pronunciation evaluation using these speech recognition technologies still has the problem that the evaluation accuracy decreases when the number of sentences for evaluation is small [6][10].

The advent of Deep Neural Networks (DNNs) improved the accuracy of pronunciation evaluation in the CAPT system. Thanks to the larger amount of training data and more complex models, the DNN-based pronunciation evaluation models effectively learn human pronunciation. However, there is still a big difference between machine and human evaluations. In machine pronunciation evaluations, the evaluation is based only on the audio signal and its transcription of the utterance. Thus, the evaluation does not take the pronunciation difficulty of the sentence into account.

On the other hand, human evaluation incorporates the difficulty of the sentence. When the sentence is easy to pronounce, humans tend to rate them higher than machines. Thus, the human evaluation score could be better than machine evaluation. If we use many sentences with various difficulties for the evaluation, this discrepancy can be canceled by averaging the evaluation scores. However, when we use only one or two utterances for evaluation, this bias makes the machine and human scores different.

We need to solve this problem because it is crucial to accurately evaluate the pronunciation when using even one utterance to reduce the time cost and physical burden on learners.

Therefore, in this paper, we improve the conventional pronunciation evaluation method by considering the pronunciation difficulty score of a sentence (sentence speakability) in addition to the scores obtained from the speech recognizer. We can improve the evaluation accuracy for a small number of utterances by integrating these scores. To this end, we first created a sentence speakability database and then examined the linguistic features for estimating the speakability. Then we investigate how speakability improves the accuracy of pronunciation evaluation.

Section 2 describes the conventional pronunciation evaluation system and how we improve the system using sentence speakability. In Section 3, we describe a database of sentence speakability. To develop the proposed system, we need the ground truth of sentence speakability for all of the sentences in the database; however, it is too costly to annotate speakability manually. Therefore, we estimate the speakability's ground truth by decomposing the human-evaluated scores (that contain both difficulties of the sentences and the speaker's speaking skill) into the speakability and the personal pronunciation ability score. In Section 4, we investigate features useful for sentence speakability estimation. In addition to the conventional features [11], we consider features that take the sentence's articulatory characteristics and the speaker's psychological characteristics into account. Then, in Section 5, we perform the speakability estimation and conduct pronunciation evaluation using the estimated speakability.

2. Automatic pronunciation evaluation

This section first describes the pronunciation evaluation system based on the DNN-HMM and then describes how to incorporate the sentence speakability. Figure 1 shows the overview of the proposed system. The proposed method adds the estimated sentence speakability in addition to the conventional pronunciation score to obtain the new score.

http://dx.doi.org/10.21437/Interspeech.2021-1132

2.1. Automatic English pronunciation evaluation using reference-free error rate

As a conventional method, we use a pronunciation evaluation system using Reference-free Error Rate (RER) [12]. A DNN-based speech recognizer trained from a non-native English speech database shows a low word error rate (WER) for non-native utterances, while the WER using a recognizer trained from native English speech depends on the proficiency of the speaker [13]. Therefore, this system introduced RER, which is the Levenshtein distance between two recognition results obtained from the recognizers developed by the native and non-native speech. This system improves the conventional pronunciation evaluation system's problem that could not evaluate speech without a reference transcription.

Let I be the number of sentences for evaluation,

$$T_{nat} = (T_1^{(nat)}, \cdots, T_I^{(nat)}) \tag{1}$$

and

$$T_{non} = (T_1^{(non)}, \cdots, T_I^{(non)}) \tag{2}$$

be the recognition results obtained from native and non-native speech recognizers, respectively. Then the RER is calculated as

$$RER = \frac{\sum_{i=1}^{I} d_L(T_i^{(non)}, T_i^{(nat)})}{\sum_{i=1}^{I} \left|T_i^{(non)}\right|} \tag{3}$$

where $d_L(X, Y)$ denotes the Levenshtein distance and $|T_i^{(non)}|$ is the number of word in the i-th sentence.

Finally, the system combines the RER and log-likelihoods obtained from the two recognizers using linear regression. In addition, we could improve the estimation accuracy by combining the recognizers' scores based on both the DNN-HMM and GMM-HMM. We denote the machine-estimated score as $\hat{S}_H(u)$ and the human score as $S_H(u)$ for utterance u.

2.2. Pronunciation evaluation considering the sentence speakability

This study proposes an improved pronunciation evaluation system that combines sentence speakability with the acoustic-based scores such as log-likelihood and RER.

First, we created a database of sentence speakability. We used sentence speakability scores $S_a(u)$ from this database as the ground truth. Then, we developed a sentence speakability estimation model and estimated the sentence speakability score from the linguistic features[11], which we describe as $\hat{S}_a(u)$. $\hat{S}_a(u)$ is combined with other features obtained from the speech recognizers to calculate the final pronunciation evaluation score $\hat{S}_H(\hat{S}_H(u), \hat{S}_a(u))$ by linear regression.

3. Development of a database of sentence speakability

To develop a sentence speakability estimation model, we created a database of $S_a(u)$ [11]. Since it is costly to annotate $S_a(u)$ manually, we automatically create it from $S_H(u)$. Non-native databases like $S_H(u)$ are often smaller than native databases, so the use of native and unlabeled databases has been revised in previous pronunciation evaluation systems [14][15]. The $S_H(u)$ database in this study also contains missing values for some sentence-speaker combinations. Thus, we propose

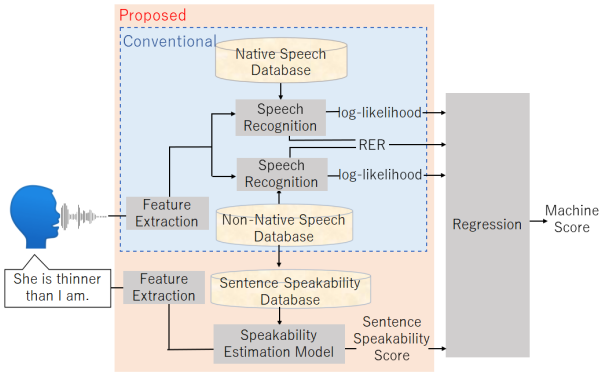

Figure 1: *Overview of the proposed pronunciation evaluation system.*

a method that automatically obtains $S_a(u)$ from $S_H(u)$ with missing values using the matrix factorization approach.

To obtain $S_a(u)$ at a low cost, we assume that we can decompose $S_H(u)$ into $S_a(u)$ and the speaker's pronunciation ability score $S_p(u)$. In this section, $S_H(u)$ is represented as a matrix $\boldsymbol{S}_H = \{S_H(i,j)\}$ that has I rows (the speakers) and J columns (the sentences). Then $S_H(u)$ is decomposed into the sum of $S_p(u)$ and $S_a(u)$ vectors $\boldsymbol{S}_p = \{S_p(i)\}$ and $\boldsymbol{S}_a = \{S_a(j)\}$.

$$S_H(i,j) = S_p(i) + S_a(j) + \epsilon(i,j) \tag{4}$$

Then we estimate $S_p(i)$ and $S_a(j)$ by minimizing the squared sum of $\epsilon(i,j)$. Since there are many missing values in $S_H(i,j)$, we introduce the mask matrix $\boldsymbol{P} = \{P(i,j)\}$ where $P(i,j) = 0$ when $S_H(i,j)$ is missing and $P(i,j) = 1$ otherwise. The squared sum of error is shown as follows.

$$E = \sum_{i,j} P(i,j)\epsilon^2(i,j) \tag{5}$$

The solutions can be calculated iteratively as follows.

$$S_p(i)^{(n+1)} = \frac{1}{N_i} \sum_j P(i,j) \left(S_H(i,j) - S_a(j)^{(n)}\right) \tag{6}$$

$$S_a(j)^{(n+1)} = \frac{1}{M_j} \sum_i P(i,j) \left(S_H(i,j) - S_p(i)^{(n)}\right) \tag{7}$$

where N_i and M_j are given as follows.

$$N_i = \sum_j P(i,j), \quad M_j = \sum_i P(i,j) \tag{8}$$

4. Linguistic features

In order to perform sentence speakability estimation from sentences, we investigated linguistic features. In the previous study [11], we found that the number of phoneme types and the word familiarity in a sentence [16][17] are helpful. In this study, we investigated the surprisal [18][19] and the phonological features [20].

4.1. Surprisal

Surprisal is a model that calculates the mental load of word processing when reading a sentence [18][19]. We calculate the surprisal of the i-th word W_i using the partial sentence generation

probability $P(W_1, \cdots, W_i)$ as follows.

$$Surprisal(W_i) = -\log \frac{P(W_1, \cdots, W_i)}{P(W_1, \cdots, W_{i-1})} \quad (9)$$

In this study, we use the incremental top-down parser [21] to extract the surprisal. This parser is a syntactic analyzer that performs phrase structure analysis and calculates generation probabilities using probabilistic context-free grammar (PCFG). In particular, we used the lexical surprisal, which is one of the parser outputs and represents the generation probability up to the word position. The maximum value of the lexical surprisal in a sentence was used as the surprisal of the sentence. For the analysis, we used a total of 49,207 sentences from chapters 0 to 24 of the Wall Street Journal in Penn Treebank [22].

4.2. Phonological features

We extracted features that represent the articulatory characteristics of sentences using the phonological features. Conventionally, phonological features have been used to evaluate phonemes, especially to evaluate the closeness of phonemes to each other [14]. In this study, we propose features that can represent articulatory difficulties for non-native speakers to pronounce. The phonological features represent articulatory styles such as vowels, affricate, and front tongues, as well as attributes of articulatory position [23][24]. We used 28 phonological features in this study as shown in Table 1.

From the phonological features of a sentence, we calculate a single value Y that is used as the feature to calculate the speakability. Y is the predicted speakability from the phonological features using the multiple linear regression, as follows. Let n be the number of phonemes of a sentence, $p_1, \cdots, p_n$ be the phonemes, and $F_k(p) \in \{0, 1\}$ be the k-th phonological feature of phoneme p. Then we calculate the probability of phonological feature and Y as follows.

$$P_k = \frac{1}{n} \sum_{i=1}^{n} F_k(p_i) \quad (10)$$

$$Y = \sum_{k=1}^{K} b_k P_k \quad (11)$$

Here, b_k is the multiple regression coefficients.

Table 1: *Phonological features*

Consonants	Voicing	Voiced, Voiceless
	Points of Articulation	Bilabial, Labiodental, Dental, Alveolar, Post-alveolar, Palatal, Velar, Glottal
	Manner of Articulation	Plosive, Nasal, Tap, Fricative, Approximant, Lateral
Vowels	Frontness	Front, Central, Back
	Height	Narrow, Semi-narrow, Semi-open, Open
	Rounding	Rounded
	Rhoticity	R-colored
	Tenseness	Tense

5. Experiments

5.1. Non-native speech database

We took $S_H(u)$ of the non-native utterances from the English Read by Japanese (ERJ) corpus [25], which includes scores for English sentences read by Japanese and Americans. The utterances have evaluation scores given by English native speakers. The total number of utterances is 29,744. The scores were rated on a 5-point scale (1 being the worst and 5 being the best) by five native English-speaking raters. In this study, we used the phoneme ratings for spoken sentences. The number of utterances was 1,900 (723 sentences, 190 speakers).

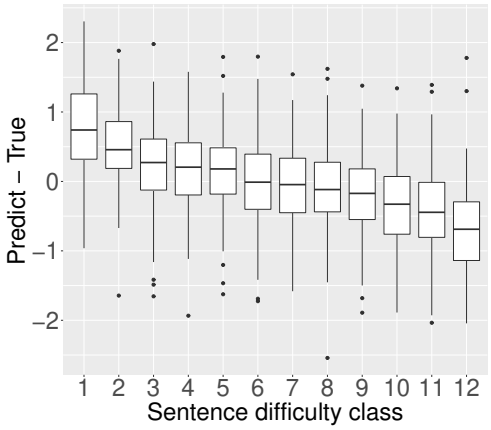

Figure 2: *Bias of predicted pronunciation scores $\hat{S}_H(u)$ with respect to speakability $S_a(u)$*

5.2. The bias of estimated scores by speakability

Figure 2 shows the difference between $\hat{S}_H(u)$ and $S_H(u)$ for each $S_a(u)$. The conditions of the speech recognizer to calculate $\hat{S}_H(u)$ are shown in Table 2. In Figure 2, the speakability scores are divided into 12 classes, with high values indicating that the sentence is easy to pronounce. This result indicates that when a sentence is easy to pronounce, the machine tends to give a lower score than the human score, and vice versa. In this study, we aim to improve the evaluation accuracy by eliminating this bias by incorporating the sentence's speakability.

5.3. Evaluation of linguistic features

In evaluating linguistic features, we calculated the correlation coefficients with $S_a(u)$ to evaluate how well the features represent the pronunciation difficulty. Here, since the weights of phonological features are obtained by multiple regression, four-fold cross-validation (181 test sentences and 523 training sentences) was conducted to evaluate phonological features. Table 3 shows the correlation coefficients between a feature and $S_a(u)$. We examined four features: two features used in the previous work [11] (the number of phoneme types and the word familiarity) and two features described in the previous section (the surprisal and the phonological feature). Pearson's correlation coefficient (PCC) was used as the correlation coefficient.

The results showed that the proposed two features had comparable correlations with the previously devised features. When the speakability was estimated using only the conventional two features, the correlation coefficient with $S_a(u)$ was 0.41 ($p < 0.01$), whereas the correlation coefficient with all features was 0.48 ($p < 0.01$). This result indicates that the inclusion of the surprisal and the phonological features helped improve the accuracy of speakability estimation.

Table 2: *Conditions of speech recognizer*

GMM-HMM	Model	AM : Triphone, LM : 3-gram
	Training data	AM : TIMIT[26] (Native, 6,300 utt.), ERJ (Non-Native, 29,744 utt.), LM : TIMIT (2,343 sent.), ERJ (980 sent.), English conversation text[27]
DNN-HMM	Model	kaldi NNET2
	Training data	Non-Native : ERJ (29,744 utt.), Native : TIMIT (6,300 utt.)
	Features	40 dim. MFCC
	Activation	ensemble p-norm[28]
	Input dim.	360
	# hidden layers	4
	p-norm dim.	Input 1000/Output 200
	Output dim.	1551

Table 3: *Correlation coefficients between a feature and speakability* $S_a(u)$

	|PCC|
Number of phoneme types	0.37
Word familiarity	0.32
Surprisal	0.29
Phonological feature	0.33

5.4. Accuracy of automatic English pronunciation evaluation

Finally, we constructed a speakability estimation system and evaluated the pronunciation evaluation system's accuracy when the sentence speakability is incorporated. The speakability estimation model is developed by multiple regression between the linguistic features and $S_a(u)$. The conditions of the speech recognizer development for automatic English pronunciation evaluation are shown in Table 2 [12]. The multiple regression calculates $\hat{S}_H(\hat{S}_H(u), \hat{S}_a(u))$ with $S_H(u)$ using the RER, log-likelihood obtained from speech recognition, and $\hat{S}_a(u)$ estimated from the linguistic features. The linguistic features include the number of phoneme types, word familiarity, surprisal, and phonological features.

For pronunciation evaluation, all sentences in ERJ (723 sentences) were used for the regression to obtain the phonological features and to build the speakability estimation system. Finally, the pronunciation evaluation accuracy was evaluated using two-fold cross-validation (950 test utterances and 950 training utterances).

For the accuracy evaluation, we measured the correlation between $\hat{S}_H(\hat{S}_H(u), \hat{S}_a(u))$ and $S_H(u)$, and $\hat{S}_H(u)$ and $S_H(u)$, respectively. $\hat{S}_H(\hat{S}_H(u), \hat{S}_a(u))$ and $\hat{S}_H(u)$ were calculated for each unit from one to ten utterances to clarify the accuracy's dependence on the utterance's length. In the case of two or more utterances was used, $S_H(u)$ were averaged. The correlation between $\hat{S}_H(\hat{S}_H(u), \hat{S}_a(u))$ and $S_H(u)$, and $\hat{S}_H(u)$ and $S_H(u)$ are shown in Figure 3.

The results show that the correlation of $\hat{S}_H(\hat{S}_H(u), \hat{S}_a(u))$ (the orange line, w/ sentence speakability) is higher than $\hat{S}_H(u)$ (the blue line, w/o sentence speakability) for a small number of utterances. This means the predicted scores' bias reduced, and the evaluation became closer to human evaluation by taking the sentence speakability into account. However, when evaluating many utterances at a time, the sentences' speakabilities are averaged so that $\hat{S}_H(u)$ becomes similar to the human score. In other words, by considering the speakability, the proposed method can evaluate pronunciation with higher accuracy than the conventional method when only one or two utterances are evaluated. In particular, for a single utterance, the correlation coefficient without speakability is 0.46, while it is 0.57 when the speakability is included.

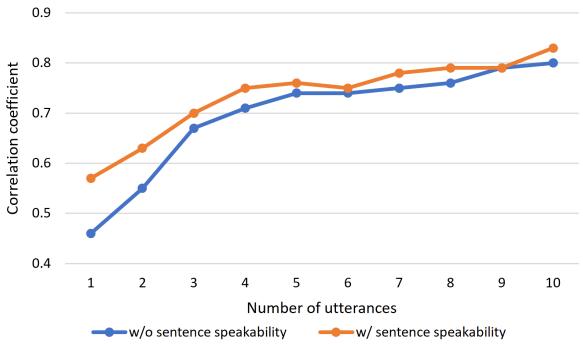

Figure 3: *Correlation between the predicted score and the human score* $S_H(u)$ *with respect to the number of utterances for evaluation*

6. Conclusions

In this study, we aimed to improve pronunciation evaluation accuracy by considering the sentence speakability in a conventional pronunciation evaluation system. To estimate the speakability, we created a speakability database and examined features that effectively predict the speakability. The surprisal and phonological features were newly proposed as sentence features, and the usefulness of these features in estimating the speakability was demonstrated. Finally, we conducted pronunciation evaluation using the estimated speakability in addition to the conventional machine scores and found that taking the pronunciation difficulty of the sentences into account improved the correlation coefficient by up to 0.11 points for a single utterance. In this study, we used linear regression to estimate the speakability, and we expect to improve the accuracy of speakability estimation by examining the estimation method.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI JP17H00823.

8. References

[1] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (CAPT) in English," *Education and Information Technologies*, vol. 24, no. 6, pp. 3731–3743, Jul. 2019.

[2] S. M. Witt and S. J. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," S. Jager, J. Nerbonne, and A. van Essen, Eds. Language Teaching and Language Technology, 1998, pp. 25–35.

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

[4] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training*, vol. 1, Jun. 2012, pp. 1–8.

[5] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, vol. 30, no. 2-3, pp. 121–130, Feb. 2000.

[6] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech & Language*, vol. 21, no. 1, pp. 219–230, Jan. 2007.

[7] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, Mar. 2015.

[8] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2016, pp. 6135–6139.

[9] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2016, pp. 6140–6144.

[10] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)." in *INTERSPEECH*, Jan. 2013, pp. 1886–1890.

[11] S. Naijo, Y. Chiba, T. Nose, and A. Ito, "Analysis and estimation of sentence speakability for English pronunciation evaluation," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, Oct. 2020, pp. 449–451.

[12] J. Fu, Y. Chiba, T. Nose, and A. Ito, "Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models," *Speech Communication*, vol. 116, pp. 86–97, Jan. 2020.

[13] ——, "Evaluation of English speech recognition for Japanese learners using DNN-based acoustic models," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Springer, Nov. 2018, pp. 93–100.

[14] V. Laborde, T. Pellegrini, L. Fontan, J. Mauclair, H. Sahraoui, and J. Farinas, "Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information," in *INTERSPEECH*, Sep. 2016, pp. 2686–2690.

[15] B. Lin, L. Wang, X. Feng, and J. Zhang, "Automatic scoring at multi-granularity for L2 pronunciation," in *INTERSPEECH*, Oct. 2020, pp. 3022–3026.

[16] L. Katz, L. Brancazio, J. Irwin, S. Katz, J. Magnuson, and D. Whalen, "What lexical decision and naming tell us about reading," *Reading and writing*, vol. 25, no. 6, pp. 1259–1282, May 2012.

[17] D. A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The English lexicon project," *Behavior Research Methods*, vol. 39, no. 3, pp. 445–459, Aug. 2007.

[18] J. Hale, "A probabilistic Earley parser as a psycholinguistic model," in *Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, vol. 2, Jun. 2001, pp. 1–8.

[19] D. M. Howcroft and V. Demberg, "Psycholinguistic models of sentence processing improve sentence readability ranking," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, Jan. 2017, pp. 958–968.

[20] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, Oct. 2000.

[21] B. Roark, A. Bachrach, C. Cardenas, and C. Pallier, "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Aug. 2009, pp. 324–333.

[22] A. Taylor, M. Marcus, and B. Santorini, "The Penn treebank: an overview," *Treebanks:Building and using parsed corpora*, vol. 20, pp. 5–22, Jan. 2003.

[23] T. Nitta, T. Onoda, M. Kimura, Y. Iribe, and K. Katsurada, "One-model speech recognition and synthesis based on articulatory movement HMMs," in *INTERSPEECH*, Sep. 2010, pp. 2970–2973.

[24] T. Nitta, S. Manosavan, Y. Iribe, K. Katsurada, R. Hayashi, and C. Zhu, "Pronunciation training by extracting articulatory movement from speech," in *Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training*, Jun. 2012, pp. 75–78.

[25] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," in *Proceedings of Intelligent Control and Automation (ICA)*, vol. 1, no. 2004, Jan. 2004, pp. 557–560.

[26] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, Aug. 1990.

[27] Spears, B. Richard A, Birner, S. Kleindler, and L. Nisset, *Conversational American English*. McGraw-Hill Education, 2010.

[28] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 215–219.