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

Lexical Density Analysis of Word Productions in Japanese English Using Acoustic Word Embeddings

Shintaro Ando, Nobuaki Minematsu, Daisuke Saito

Graduate School of Engineering, The University of Tokyo, Japan
{s_ando,mine,dsk_saito}@gavo.t.u-tokyo.ac.jp

Abstract

In L2 pronunciation, what kind of phonetic errors are more influential to intelligibility reduction? Teachers say that learners' utterances become unintelligible when words are pronounced with such errors that make the words misidentified as others. In this paper, we focus on Japanese English (JE), where the number of phonemes of the L1 (Japanese) is much smaller than that of the L2 (American English, AE). Since learners often substitute L1 phonemes when speaking in L2, some words are expected to be pronounced not distinctively enough in JE, which may result in word misidentification. This implies that words of JE will exist phonetically closer to each other in a space where words are distributed. In this paper, lexical density analysis of JE and AE is carried out using acoustic word embeddings. Word productions in JE and AE, extracted from the ERJ corpus, are mapped as points in an acoustic word embedding space obtained by network training with the WSJ corpus. Experiments show that significantly higher density is found in JE than in AE and it is also found in poor learners than in good learners.
Index Terms: lexical density, L2 pronunciation, acoustic word embedding, proficiency, CALL

1. Introduction

Many language teachers claim that the practical goal of pronunciation training is to acquire an intelligible enough pronunciation, not a native-like pronunciation [1–3]. Possible reasons of intelligibility reduction are related to semantics, pragmatics, syntax, lexical choice as well as pronunciation in L2 utterances [4]. In this paper, we focus on pronunciation, especially intelligibility reduction caused by reduced distinctiveness between phonetic realizations of different phonemes. If two phonemes /x/ and /y/ are pronounced phonetically in the same way, a minimal pair of words, which differ only by one phoneme, /x/ in a word and /y/ in the other, cannot be distinguished by listeners. In [5, 6], researchers attempted to rank phonemic contrasts according to their importance in distinguishing English words in terms of pronunciation. For example, many English word pairs are distinguished by the /l/-/n/ contrast while the /ð/-/d/ contrast distinguishes relatively fewer pairs of words. To discuss differences of importance, researchers introduced a notion of functional load (FL) and they described the /l/-/n/ contrast as higher FL and the /ð/-/d/ contrast as lower FL.

The functional load principle was initially discussed in L1 English and then, it was transferred and used in analyzing L2 English. In [6], it was discussed whether L2 utterances including phonemic errors of higher FL reduced subjective comprehensibility scores than those of lower FL. Here, FL was defined as a list of phonemic contrasts that are ranked based on their communicative value. These contrasts were developed by taking into account minimal pairs of frequent words, the degree of neutralization among regional dialects, and the segmental position within a word. Recently, the pedagogical value of the functional load principle was revisited in [7] using Japanese English utterances, and the impact of high FL errors on perceived comprehensibility was re-investigated. Based on these empirical studies, teachers can decide which phoneme replacement errors in actual L2 utterances should be pointed out in class.

With reduced distinctiveness between phonemes, it is easily expected that distinctiveness among words will be reduced. With the phonemic contrasts of higher FL, derived by the FL principle, it is possible to assess intelligibility or distinctiveness of the pronunciation of a specific learner by detecting actual phonetic errors in his/her utterances. With speech technologies, however, distinctiveness of words produced in a non-native pronunciation can be quantitatively evaluated more directly without the list of phonemic contrasts with higher FL. In [8, 9], phoneme HMMs were trained separately for AE with the Wall Street Journal (WSJ) corpus and for JE with the English Read by Japanese (ERJ) corpus. With the two kinds of HMMs, every word in the CMU dictionary [10] is formed in two ways as a sequence of AE HMMs and that of JE HMMs. Separately in AE and JE, word-to-word distance was calculated as Dynamic Time Warping (DTW) distance between the corresponding two HMM sequences. With N being the vocabulary size, the $N \times N$ distance matrix was obtained separately for AE and JE. With this matrix, the lexical density of each word was calculated, which means the number of different words located at a distance shorter than a threshold to that specific word. Experiments showed that JE had higher lexical density, meaning higher confusability or lower distinctiveness than AE. However, this analysis was done holistically, not for individual learners.

In the era of network-based speech technology, a more feasible technical framework for lexical density analysis is available, which is acoustic word embedding (AWE) [11–13]. In this framework, any input spoken word is converted into its embedding vector, which is just a point in the word embedding space. Different from the above works [8, 9], word-to-word distance is calculated simply as point-to-point distance such as the Euclidian distance or cosine distance, not as DTW distance between HMM sequences. Further, individual instances of spoken words can be directly used for analysis. In [8, 9], JE HMMs and AE HMMs were trained with the ERJ corpus and the WSJ corpus, and only a holistic comparison was made between JE and AE. With the word embedding space, however, any single spoken word can be mapped into the space and the number of competing words can be calculated as the number of words located at a shorter distance to that specific spoken word.

To the best of our knowledge, this paper conducts lexical density analysis on L2 English for the first time using AWEs. The aim of this study is to show possibility of AWE for this purpose, not to compare AWE-based analysis and HMM-based analysis because the latter analysis takes computation time and can make only a holistic comparison between JE and AE.

http://dx.doi.org/10.21437/Interspeech.2021-853

Table 1: *The number of phonemes in English and Japanese*

	vowels	consonants
American English	16	24
Japanese	5	19

Table 2: *Typical phoneme substitutions observed in JE*

ID	substitution rules	contextual conditions
C1	r → l	^–r+*, (C)–r+*, (V)–r+(V)
C2	θ → s	
C3	ð → z	
C4	v → b	
V1	oʊ → ɔː	
V2	æ → ʌ	
V3	əːr → ɑː	
V4	ʌ → ɑː	can be applied prior to V2
V5	r → ɑː	(D)–r+(C), (D)–r+$

C: consonant, V: vowel, D: diphthong

2. Phonemic characteristics of American English, Japanese, and Japanese English

Some characteristics of the phonological systems of AE, JE, and Japanese (JP) are explained [14, 15]. A part of characteristics of JE are attributed to those of its L1. Table 1 shows the number of vowels and consonants in AE and JP. Clearly shown, those numbers are smaller in JP, especially the number of vowels is less than a third of that in AE. To acquire a good pronunciation of English, Japanese learners have to learn how to produce new phonemes, which are unseen in JP. As well known as L1 transfer, learners often substitute L1 phonemes for those new L2 phonemes. Taking into account such behaviors of learners and Table 1, it is easily expected that distinctiveness of the JE phonemes will be lower compared to that of the AE phonemes. It is reasonable to expect that JE word productions will exhibit higher density compared to AE word productions.

What kind of inadequate phonetic realizations of phonemes are observed actually in JE? [16] analyzed JE utterances and summarized typical and salient substitutions as well as specific contexts where those substitutions are found. Table 2 shows its summary. Cn are consonant substitutions. In JP, /r/ and /l/ are not distinguished and these two are perceived as the same phonemic class. /θ/, /ð/, and /v/ are not found in JP and /s/, /z/, and /b/ are substituted instead, respectively. Vn are vowel substitutions. Since JP has only five vowels, several different AE vowels are merged into one vowel class. In Section 4, we will show experimental results of lexical density analysis of JE word productions as well as some results are also shown by using synthetic word productions generated by an AE speech synthesizer with phonemic substitutions applied based on Table 2. It should be noted that this table shows only typical and salient phoneme substitutions in JE. In actual JE utterances, however, some additional characteristics can be easily found. For example, since the unit of speech production of JP is mora, which is open syllable in the form of V and CV, vowel epenthesis is found not rarely, especially in utterances given from beginning learners.

3. Speech corpus and technology used for lexical density analysis

3.1. English Read by Japanese (ERJ) corpus

For lexical density analysis for JE, we extracted spoken word samples from the ERJ corpus [17–19]. This corpus is divided into a sentence set and a word set, either of which includes a phonemically-balanced subset. The total number of words of the phonemically-balanced word set is 250, which is divided equally into five subsets (W1 to W5). The ERJ corpus has 100 male and 100 female college students as non-native speakers, and each of them read aloud one word subset. The corpus also has 20 native speakers as model speakers, each of whom read aloud a half of the phonemically-balanced word set.

In the corpus, manual scores are assigned to some word samples and sentence samples of each student. The raters are teachers of English who had good experiences of teaching English pronunciation to Japanese students. Some different strategies were adopted in scoring such as phonetic realization of phonemes, prosodic realization of phrase intonation and word stress, etc. For lexical density analysis, we will refer to the manual scores rated in terms of phonetic realization of phonemes.

3.2. Acoustic Word Embedding (AWE)

AWE is a method of representing a spoken word sample as a vector of a fixed dimension d [11, 12]. Here, the word sample is embedded as a point (vector) in the d-dimensional space. The vector is often called as embedding and the space is called as embedding space. Mapping from word productions to AWEs is trained so that phonetically similar words are placed closer to each other in the embedding space, and dissimilar words are placed at a longer distance to each other there. AWE was originally proposed to improve the performance of automatic speech recognition, and it was also used in the task of spoken term detection [13, 20]. More recently, AWE spaces were analyzed and tested whether they can simulate the mental lexicon of humans, especially it was discussed whether phonetic learning was realized in the spaces in a similar way to infants' phonetic learning [21, 22]. In the present study, we assume the AWE space trained only with the WSJ corpus, which contains only AE read-aloud sentences, as the mental lexicon of AE speakers.

In this embedding space, we plot AE and JE word productions extracted from the ERJ corpus. Distributional differences, i.e. lexical density, between AE and JE are analyzed. Out of various technical implementations of AWE, in this study, we use the Multi-View Siamese (MVS) AWE model [23] and the Corresponding AutoEncoder (CAE) AWE model [24].

3.2.1. AWE based on Siamese network

The Siamese network was proposed as a method of metric learning [25], which takes two samples as input and generates two embedding vectors in the embedding space. The network was trained so that similar samples are plotted closer to each other in the space and dissimilar samples are plotted more distant to each other. In [13], the internal representation at the final and hidden layer of RNN was adopted as AWE and a better performance of Spoken Term Detection (STD) was attained. For input sequence of acoustic observations $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$, a positive sample of the same class $\boldsymbol{X}^+$, and a negative one of a different class $\boldsymbol{X}^-$, the following triplet loss was used.

$$L_{\text{triplet}}(\boldsymbol{X}, \boldsymbol{X}^+, \boldsymbol{X}^-) = \max\left\{0, d\big(f(\boldsymbol{X}), f(\boldsymbol{X}^+)\big) + m - d\big(f(\boldsymbol{X}), f(\boldsymbol{X}^-)\big)\right\} \quad (1)$$

where m is a hyper parameter and, when it becomes larger, $d\big(f(\boldsymbol{X}), f(\boldsymbol{X}^-)\big)$ tends to have a larger value.

In the present study, out of several technical proposals of the Siamese-based AWEs, the Multi-View Siamese (MVS) model [23] is used, which showed a higher STD performance compared to single-view approaches. In this model, in addition to acoustic observations $\boldsymbol{X}$, its corresponding visual observations,

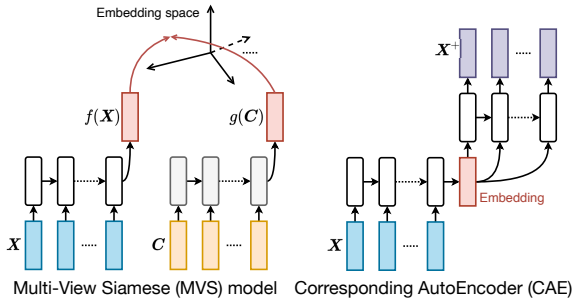


Figure 1: *Acoustic word embedding based on MVS and CAE*

i.e. graphemes, $\boldsymbol{C} = \boldsymbol{c}_1, \boldsymbol{c}_2, \ldots \boldsymbol{c}_L$, are also used to build a shared embedding space among the two modalities, acoustic and visual. By using the following integrated loss, $\boldsymbol{X}$'s embedding, $f(\boldsymbol{X})$, and $\boldsymbol{C}$'s embedding, $g(\boldsymbol{C})$, will be plotted closer in the shared embedding space. It should be noted that, unlike the triplet loss, distances between two embeddings in the same modality are not used in L_{MVS}. In [23], it was empirically shown that a higher STD performance was obtained without the distances in the same modality, which implies that use of multiple modalities in the shared space is very effective. Figure 1 illustrates this Multi-View Siamese (MVS) AWE model.

$$\begin{aligned} &L_{\mathrm{MVS}}(\boldsymbol{X}, \boldsymbol{X}^+, \boldsymbol{X}^-, \boldsymbol{C}, \boldsymbol{C}^+, \boldsymbol{C}^-) \\ &= \max\left\{0, d(f(\boldsymbol{X}), g(\boldsymbol{C})) + m - d\big(f(\boldsymbol{X}^+), g(\boldsymbol{C}^-)\big)\right\} \\ &+ \max\left\{0, d(f(\boldsymbol{X}), g(\boldsymbol{C})) + m - d\big(f(\boldsymbol{X}^-), g(\boldsymbol{C}^+)\big)\right\} \quad (2) \end{aligned}$$

3.2.2. AWE based on Correspondence AutoEncoder (CAE)

A different training strategy was proposed in [24]. Here, unsupervised spoken term discovery was applied to a speech corpus without annotation, and discovered segments for a specific word query were treated as samples of that word class. For a pair of segments of the same class $(\boldsymbol{X}, \boldsymbol{X}^+)$, the final representation obtained in an encode-decoder model was used as AWE, shown in Figure 1. By inputing the AWE to the decoder, it can generate such $\boldsymbol{X}^+$ that minimizes the following loss. CAE does not use any negative samples and therefore distinctiveness among different words will be reduced compared to the MVS model.

$$L_{\mathrm{CAE}}\big(\boldsymbol{X}, \boldsymbol{X}^+\big) = \sum_{t=1}^{T^+} ||\boldsymbol{x}_t^+ - f_t(\boldsymbol{X})||^2 \quad (3)$$

Following [21, 22], we may dare to regard the MVS-based embedding space as the mental lexical space of AE adults because graphemic representations and negative samples are used very effectively. On the other hand, the CAE model uses only positive acoustic samples even in an unsupervised way. The CAE-based space can be regarded as the mental lexical space of AE speakers with immature language knowledge, which may correspond to younger AE kids.

4. Experiments

4.1. Training of the two models with the WSJ corpus

The two AWE models of MVS and CAE were trained with the WSJ corpus, where word alignments were obtained by using the WSJ-Kaldi nnet3 recipe [26]. To build the two models, the WSJ corpus was divided into 18:1:1 parts for training, development, and testing. For each part, detected word segments which were longer than 0.1 sec and shorter than 2.0 sec were used. For each word in the training part, only one instance was extracted from each speaker. This resulted in that the maximum number of instances per word was 50. The vocabulary size of the training part was 12,018 and the total number of word segments was 187,458, which were so long as about 26 hours. From those utterances, 13-dimensional MFCCs were extracted with their Δ and $\Delta\Delta$ to train the two AWE models.

For the MVS model, both acoustic-view and character-view models used 512-dimensional bidirectional two-layered LSTMs. The values of their hyper-parameters were set by following [23]. The embedding vector was derived by connecting the final frames of the hidden layers of the BLSTMs.

For the CAE model, word fragments were detected in a supervised way as in the MVS model. The encoder and the decoder were both trained as 256-dimensional unidirectional three-layered GRUs. The batch size was 128 and Adam optimization was performed with learning rate of 0.001.

For either model of MVS and CAE, after each epoch of training, both the models were verified using the development data. The highest precision was obtained after 300 epochs for MVS and 100 epochs for CAE, which were adopted as the final models actually used for lexical density analysis.

4.2. Lexical density calculated for the individual speakers

The ERJ corpus has phonemically-balanced word sets (W1 to W5), each of which contains 50 words. The number of learners per word set is about 40 and that of native speakers is 8 or 9. By using the MVS model or the CAE model, each word segment was converted to its embedding vector, and each speaker had his/her 50 word embeddings. For them, agglomerative hierarchical clustering was applied. Here, the group average method was used and the distance between two embeddings was calculated as their cosine distance c (c=1−cos ϕ, 0≤c≤2).

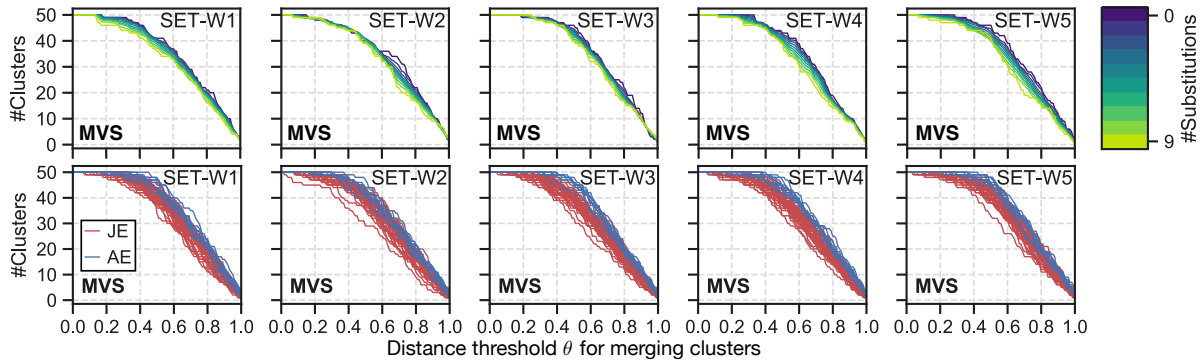
By controlling a distance threshold θ for merging, 50 embeddings are clustered into k clusters, where 1≤k≤50. If the distance between two samples is smaller than θ, they should be merged. In other words, if θ=0, then k=50, and if θ=2, then k=1. If we set θ to an adequate value and two speakers show different numbers of clusters, we can say that the speaker with the higher number of clusters has a more distinctive pronunciation than the other speaker.

Table 2 shows typical and salient phoneme substitutions found actually in JE. In the experiment, we also carried out lexical density analysis using JE word productions *simulated* based on this table. We used a Tacotron2-based AE speech synthesizer [27] and had it read aloud 250 sequences of phonemes, corresponding to W1 to W5. Table 2 defines nine substitutions, and when m substitutions are selected, $\binom{9}{m}$ combinations of m substitutions are possible. By applying every possible combination to subset Wn, we calculated the averaged number of clusters of Wn when m substitutions were applied.

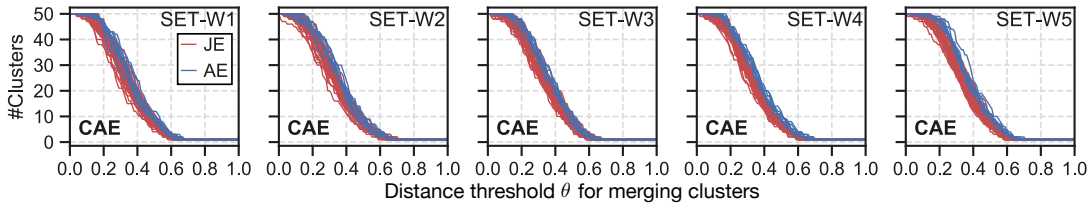
4.3. Results and discussion

4.3.1. Simulated productions vs. actual productions

The numbers of clusters obtained by agglomerative hierarchical clustering are plotted in Figure 2, where the x-axis represents threshold θ. Figure 2a compares the results between simulated productions and actual ones, calculated with the MVS model. Probably because each word subset is a phonemically-balanced subset, similar curves are found among the subsets. In



(a) *The number of clusters in simulated productions (top) and actual productions (bottom), calculated with the MVS model*



(b) *The number of clusters in actual productions, calculated with the CAE model*

Figure 2: *The number of clusters in simulated and actual productions, calculated with the two AWE models*

each word subset of the simulated productions, 10 curves are drawn, which correspond to m ($0 \leq m \leq 9$). Here, 0 means no substitution and 9 indicates that all the substitutions in Table 2 were applied. If we use any fixed value for θ, we can say that the more substitutions are applied (lighter green), the smaller number of clusters is found, indicating higher lexical density and lower distinctiveness. In the results of the actual productions, the number of curves in a word subset corresponds to the number of students and native speakers in the word subset. Although several students attained similar density to AE, a majority of them had lower numbers of clusters, indicating higher density. When we compare the number of clusters in the simulated productions and that in the actual ones, a good similarity is found between the curves with m=0 and the AE curves. It is clear, however, that some students' actual curves are lower than the simulated curves with m=9. This implies that, in actual productions, some other types of pronunciation deviations than phoneme substitutions are easily found, and that the functional load principle, where only phoneme substitutions are considered, may need some theoretical refinements to better evaluate the communicative value of actual L2 pronunciations.

4.3.2. *MVS embeddings vs. CAE embeddings*

The results obtained with the CAE model and the actual productions are plotted in Figure 2b. When we compare the results of both models within the same word subset, the MVS model always shows larger numbers of clusters. This indicates that the MVS model evaluates a given set of word productions as more distinctive than the CAE model. This is reasonable because the MVS model was trained with negative samples and graphemic representations, and it will enhance distinctiveness or contrast between different words. Which model should be used pedagogically? This question is discussed in the following section.

4.3.3. *Good learners vs. poor learners*

Since each word subset had about 40 learners, from each subset, we extracted the top 10 students and the bottom 10 students by referring to their scores rated in terms of phonetic realization of phonemes in the ERJ corpus. T-test showed that, for any word subset, a significant difference was found in their scores between the two groups. We examined how AE, good learners (GL), and poor learners (PL) were evaluated in terms of the number of clusters with an adequate value for θ. Table 3 shows the results of comparison among the three groups of AE, GL, and PL, where three levels of differences are indicated as $*$, $\star$, and $\circ$. By comparing MVS and CAE in Figure 2, the numbers of clusters are decreased with CAE. In Table 3, distinctiveness between GL and PL is evaluated within each model, and distinctiveness is reduced again with CAE. Although the scores of phonetic realization in the ERJ corpus cannot be fully characterized by the numbers of clusters in Figure 2[1], it would be better to use the MVS model for practical analysis of L2 productions.

Table 3: *Significant differences of the number of clusters among AE, GL, and PL with the two models of MVS and CAE*

		W1		W2		W3		W4		W5	
		GL	PL	GL	PL	GL	PL	GL	PL	GL	PL
MVS	AE	–	$*$	$\circ$	$*$	$*$	$*$	$\star$	$*$	$\star$	$*$
(θ=0.6)	GL		$\star$		$\circ$		–		$\circ$		$\circ$
CAE	AE	–	$\circ$	–	$\star$	$\star$	–	$\circ$	$*$	$\star$	$*$
(θ=0.3)	GL		–		–		–		–		–

$*$: $p<0.01$, $\star$: $p<0.05$, $\circ$: $p<0.1$, –: $p \geq 0.1$

5. Conclusions

In this paper, acoustic word embedding techniques were examined for the first time to analyze the lexical density in JE word productions. Two models of MVS and CAE were tested and the former was shown experimentally to be more valid pedagogically than the latter. Further, JE was simulated as AE with various phoneme substitutions, but the simulated JE showed lower lexical density, compared to actual JE productions. This implies that distinctiveness of L2 pronunciations should be discussed by taking other types of deviations well into account.

[1]The ERJ scores are assigned by teachers' checking whether the individual phonemes are realized phonetically as adequate, but the numbers of clusters in Figure 2 only indicate distinctiveness among word productions of the individual speakers.

6. References

[1] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[2] J. M. Murphy, "Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching," *System*, vol. 42, pp. 258–269, 2014.

[3] J. Levis, "Revisiting the intelligibility and nativeness principles," *Journal of Second Language Pronunciation*, vol. 6, no. 3, pp. 310–328, 2020.

[4] J. Bernstein, "Objective measurement of intelligibility," in *Proc. ICPhS*, 2003, pp. 1581–1584.

[5] A. Brown, "Functional load and the teaching of pronunciation," *TESOL Quarterly*, vol. 22, pp. 593–606, 1988.

[6] M. J. Munro and T. M. Derwing, "The functional load principle in ESL pronunciation instruction: An exploratory study," *System*, vol. 34, pp. 520–531, 2006.

[7] Y. Suzukida and K. Saito, "Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle," *Language Teaching Research*, 2019.

[8] N. Minematsu, G. Kurata, and K. Hirose, "Corpus-based analysis of production and perception of Japanese English in view of the entire phonemic system of English," in *Proc. ICPhS*, 2003, pp. 1569–1572.

[9] N. Minematsu, "Pronunciation assessment based upon the compatibility between a learner's pronunciation structure and the target language's lexical structure," in *Proc. INTERSPEECH*, 2004.

[10] "The CMU Pronouncing Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[11] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 410–415.

[12] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proc. INTERSPEECH*, 2014, pp. 1053–1057.

[13] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," in *Proc. INTERSPEECH*, 2017, pp. 2874–2878.

[14] P. Roach, *English Phonetics and Phonology*. Cambridge University Press, 2009.

[15] H. Kubozono, *Handbook of Japanese Phonetics and Phonology*. De Gruyter Mouton, 2015.

[16] M. Swan and B. Smith, *Learner English –A teacher's guide to interference and other problems–*. Cambridge University Press, 2001.

[17] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "English speech database read by Japanese learners for CALL system development," in *Proc. LREC*, 2002.

[18] N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ database," in *Proc. INTERSPEECH*, 2011, pp. 1481–1484.

[19] T. Makino and R. Aoki, "English read by Japanese phonetic corpus: An interim report," *Research in Langauge*, vol. 10, no. 1, pp. 79–95, 2012.

[20] Y. Chen, S. Huang, C. Shen, H. Lee, and L. Lee, "Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval," in *Proc. IEEE Spoken Language Technology Workshop*, 2018, pp. 941–948.

[21] Y. Matusevych, H. Kamper, and S. Goldwater, "Analyzing autoencoder-based acoustic word embeddings," in *Proc. Bridging AI and Cognitive Science workshop at ICLR*, 2020.

[22] Y. Matusevych, T. Schatz, H. Kamper, N. H. Feldman, and S. Goldwater, "Evaluating computational models of infant phonetic learning across languages," in *Proc. Meeting of the Cognitive Science Society*, 2020, pp. 571–577.

[23] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," in *Proc. ICLR*, 2017.

[24] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *Proc. ICASSP*, 2019.

[25] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Advances in Neural Information Processing Systems*, J. Cowan, G. Tesauro, and J. Alspector, Eds., vol. 6. Morgan-Kaufmann, 1994.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[27] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.