9th Workshop on Speech and Language Technology in Education (SLaTE)
18-20 August 2023, Dublin, Ireland

Sensitivity to Phonemic Contrasts and Insensitivity to Non-phonemic Contrasts of Various Speech Representations Tested for L2 Speech Assessment

Haitong Sun, Yingxiang Gao, Yusuke Shozui, Tong Ma, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo, Japan
{sunhaitong,gyx,shozui,matongwithdsy,mine}@gavo.t.u-tokyo.ac.jp

Abstract

To assess the segmental aspect of L2 speech produced by various types of learners, researchers and teachers need speech representations which satisfy two conditions of being able to capture phonemic contrasts accurately and ignore non-phonemic contrasts adequately. Acoustically, both of the contrasts can be equally characterized by spectrum envelopes. Therefore, purely acoustic representations such as MFCC cannot satisfy the two conditions. Recently, phonetic posteriorgrams, which are estimated by DNN-based acoustic models of ASR, are used for L2 assessment. More recently, various kinds of self-supervised representations are proposed such as wav2vec2 and WavLM. In this study, by setting up a simple and adequate metric to examine sensitivity to phonemic contrasts and insensitivity to non-phonemic contrasts, various pretrained models are compared. Experiments show WavLM is superior to other self-supervised representations and even better than supervised representations in some cases.

Index Terms: L2 speech assessment, phonemic contrasts, (in)sensitivity, posteriorgram, self-supervised representations

1. Introduction

In the field of Computer Aided Language Learning (CALL), automatic detection of segmental errors in L2 speech and automatic scoring of the L2 speech have been discussed technically for decades [1–3]. When only an L2 speech was taken as input to a CALL system with no reference speech given, the system has to recognize what was said and evaluate how it was said by the learner. To make the system more reliable pedagogically, the L2 speech as well as its model speech are given to the system, which compares them and detects mismatched segments in the L2 speech. In language class, learners often repeat model speech, and comparison-based automatic assessment [4–9] is regarded as one of the most fundamental functions that should be realized in a pedagogically valid way.

An L2 speech and its model speech are always produced from different speakers, and in most of the cases, learners are younger than teachers. This easily means that comparison of the two speech samples should be done by using speech representations that are invariant to speaker diversity including age and gender gaps. In classical studies of comparison-based L2 speech assessment, gender-specific MFCC-based DTW was conducted, but it was very reasonable that the mismatch problem was not solved sufficiently. With the help of DNN-HMM ASR framework, phonetic posteriorgram, senone-based posterior probability $P(s_n|o_t)$ is used to represent each frame of an input speech, where s_n is senone of class n ($1 \leq n \leq N$) and o_t is speech frame at time t. Using $P(s_n|o_t)$, each frame is represented as N-dimensional vector [10], and any spectrogram is converted into its phonetic posteriorgram (PPG). In [6–9], PPG-DTW was used as a technique for comparison-based L2 speech assessment, and it showed a good comparison performance, while suppressing the unwanted but inevitable variation explained above. Generally speaking, N is so large as several thousands, and it can be reduced to a smaller dimension based on agglomerative clustering [11]. If the dimension is still large enough, however, the entire space of senones of a language may be overlapped well with that of another language. Therefore, in [7], L2 English speech samples were converted into English posteriorgrams and Japanese posteriorgrams, and PPG-DTW was examined by using the former and the latter separately. In both cases, PPG-DTW worked well while suppressing the unwanted variation.

$P(s_n|o_t)$ is often calculated using DNN acoustic models [12], which are trained in a supervised way using phonemic labels assigned to every utterance in the training data. Recently, gigantic unlabelled data have been exploited to train self-supervised models to introduce new speech representations [13–16]. Generally speaking, there are three approaches to training self-supervised models: generative, contrastive, and predictive approaches. In the generative approach, some parts of inputs are masked, and the models are trained to generate the masked segments from their unmasked surrounding contexts. The models are said to learn internal and intrinsic speech structure to restore the masked segments. In the contrastive approach, a latent space representation is trained in such a way that distance should be minimized between an anchor and its positive examples and maximized between an anchor and its negative examples. In the predictive approach, pseudo-labels are calculated in the training phrase, and the models are trained so that they can predict the labels precisely. This approach can be viewed as semi-supervised approach.

The new representations are generally obtained as outputs of a hidden layer of the models, which were used as input to various tasks [14–16]. The task performance was calculated by the representations integrated to the succeeding networks often specific to the individual downstream tasks. In this evaluation framework, however, characteristics of the new representations themselves are still unclear, especially their (in)sensitivity to phonemic and non-phonemic contrasts is unknown.

In this paper, in order to investigate how effectively these self-supervised representations work in the task of comparison-based assessment, we introduce a new and simple metric using minimal pairs of words spoken by different speakers. In pronunciation training, a minimal pair of words, which differ only by one phoneme such as “sail” (/s ey l/) and “sell” (/s eh l/), are often used. The desired representation should be very sensitive to the phonemic contrast between the two words even produced by the same speaker, and the representation should be very insen-

10.21437/SLaTE.2023-7

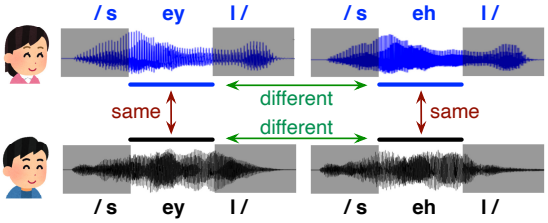

Figure 1: *Good representations should be sensitive to phonemic contrasts and insensitive to speaker contrasts.*

Table 1: *Vowel and consonant contrasts in English (left & middle) and consonant contrasts in Japanese (right)*

word pair	contrast
bap-bep	ae-eh
chick-check	ie-eh
sail-sell	ey-eh
lack-luck	ae-eh
stuck-stock	ah-aa
meat-mitt	iy-ih
luke-look	uw-uh
strike-stroke	ay-ow

word pair	contrast
bat-pat	b-p
dad-tad	d-t
git-kit	g-k
sag-thag	s-th
shap-sap	sh-s
fed-head	f-h
bat-vat	b-v
chet-jet	ch-jh
map-nap	m-n
legion-region	l-r
vet-wet	v-w

word pair	contrast
ビン-ピン	b-p
釘-国	g-n
息-位置	k-ch
逆-百	g-h
逆-客	g-k
百-客	h-k
全員-船員	z-s
白檀-爆弾	h-b
派手-羽根	d-n
派手-晴れ	d-r
:	:

sitive to the speaker-based acoustic contrast even between two "sail"s produced by two speakers, and between two "sell"s produced by two speakers. Figure 1 illustrates the desired speech representation for comparison-based L2 assessment. When the phonemic contrast is small acoustically and the speaker contrast is large acoustically, it will be difficult to exhibit high sensitivity to the phonemic contrast and high insensitivity to the speaker contrast at the same time, because both contrasts are acoustically characterized often by the same speech feature.

2. Metric adopted to compare various speech representations

The ERJ (English Read by Japanese) corpus and the JRF (Japanese Read by Foreigners) corpus include minimal pairs of words produced by native speakers and learners [17, 18]. In this study, only native productions are used[1]. Table 1 shows the English minimal pairs of words and a part of the Japanese minimal pairs of words. While both eight vowel and eleven consonant contrasts are available in English, only eighteen consonant contrasts are available in Japanese. In JRF, all the vowel-based minimal pairs are not with vowel contrasts but with durational contrasts, short vs. long. These are not used in this paper.

Let v and w be a minimal pair of word productions and $c(v)$ and $c(w)$ be the phonemic contrast segment in v and w. We have a male speaker m and a female speaker f, both of whom produced v and w, as shown in Figure 1. When we test speech representation r, after forced alignment is performed on the four word productions of v^f, w^f, v^m, and w^m, the following four kinds of DTW are conducted. $\mathrm{DTW}_r(c(v^f), c(w^f))$ and $\mathrm{DTW}_r(c(v^m), c(w^m))$ are within-speaker DTW between $c(v)$ and $c(w)$. $\mathrm{DTW}_r(c(v^f), c(v^m))$ and $\mathrm{DTW}_r(c(w^f), c(w^m))$ are cross-speaker and cross-gender DTW between two segments of the same phoneme. For v and w, we introduce the following sensitivity and insensitivity metric SI as

$$\mathrm{SI}_r(f,m) = \frac{\mathrm{DTW}_r(c(v^f), c(w^f)) + \mathrm{DTW}_r(c(v^m), c(w^m))}{\mathrm{DTW}_r(c(v^f), c(v^m)) + \mathrm{DTW}_r(c(w^f), c(w^m))}.$$

In this equation, DTW represents the averaged distance along the alignment path. Since ERJ and JRF have multiple male speakers and multiple female speakers, we can form multiple speaker pairs with different genders. After calculating SI_r for each of the pairs, we can get their average for v and w, which is denoted as $\overline{\mathrm{SI}_r}$. In the above equation, the numerator should be larger (sensitive) and the denominator should be smaller (insensitive). This means that the larger $\overline{\mathrm{SI}_r}$ is, the better we can claim that r is. If $\overline{\mathrm{SI}_r}$ is lower than 1.0, however, we can say that r is more sensitive to speaker contrast than phonemic contrast.

The proposed metric of SI is different in nature from the metrics used in previous studies [13–16]. Their metrics, or their downstream tasks, are phoneme recognition, speaker identification, emotion recognition, automatic speech recognition, query-by-example, automatic speaker verification, speaker diarization, source separation, speech enhancement, and speech translation. As mentioned in Section 1, the performances in these tasks inevitably depend on the networks that take r as input, which are generally trained discriminatively in a supervised way using labels. On the other hand, SI assesses r with no post-processing or labels, and it directly focuses on (in)sensitivity to phoneme contrasts and speaker contrasts. Phoneme recognition, one of the downstream tasks, also has to satisfy the two conditions to achieve a high performance, but it is trained with labels.

3. Speech representations tested in the experiments

In this paper, we use three kinds of speech representations: 1) raw acoustic features, 2) supervised representations of PPG, and 3) various types of unsupervised representations. In addition to insensitivity, or generalizability, over speakers, we will also test that over languages (English vs. Japanese).

3.1. Raw acoustic features

FBANK and MFCC are used as raw acoustic features. Since both phonemic and speaker contrasts are characterized largely by the spectrum envelope, $\overline{\mathrm{SI}_{\mathrm{FBANK}}}$ and $\overline{\mathrm{SI}_{\mathrm{MFCC}}}$ are expected to show lower values compared to the other representations.

3.2. Phonetic PosteriorGram (PPG)

American English PPG (AE-PPG) and Japanese PPG (JP-PPG) are used as reference representations, which are often used for comparison-based speech assessment for L2 English and L2 Japanese [6–9]. KALDI-based ASR frontend [12] is used to convert any input speech to AE-PPG and JP-PPG. The models are trained with the corpora of WSJ [19] and CSJ [20].

3.3. Self-supervised representations

3.3.1. Wav2vec2 and its cross-language version [21, 22]

Wav2vec2.0 is a self-supervised model that attempts to predict the masked feature by solving a contrastive task over quantized latent representations. The model is trained to reduce the contrastive loss between the Transformer representation and the quantized latent representation for masked segments. In the experiment, we use the Wav2vec2 Large model (W2V-L) and its

[1] In CALL studies, for validation purposes, native productions of /x/ are often used as speech samples of incorrect productions of /y/.

cross-language version (W2V-XL). The former is trained with 960 hours of LibriSpeech (English) and the latter is trained with 56k hours of 53 languages. In this paper, only W2V-XL is a multi-language model.

3.3.2. Hidden-Unit BERT [23]

Hidden-Unit BERT (HuBERT) is a self-supervised BERT-like model, which consists of a convolutional waveform encoder, a BERT encoder, a projection layer, and a code embedding layer. The model is trained to reduce the loss function between the latent representation and the pseudo-label generated internally from masked segments. In the experiment, HuBERT-large-ll60k is used, which is pretrained on 60k hours of Libri-Light English audio.

3.3.3. WavLM [24]

As a variant of HuBERT, WavLM contains a convolutional encoder and a Transformer encoder. WavLM employs gated relative position bias in the Transformer encoder to introduce relative position into the attention mechanism to better model local information. In the experiment, we use WavLM Base model (WavLM-B) and WavLM Large model (WavLM-L). The former is pretrained on 960h of LibriSpeech audio, and the latter is on 94k large-scale diverse data including 60k LibriLight audio, 10k GigaSpeech audio, and 24k VoxPopuli English audio.

3.3.4. Data2vec [25]

Data2vec is a general framework for self-supervised learning in speech, vision and language. The system uses the standard Transformer architecture with a modality-specific encoding of the input data. The training target is to predict the original unmasked training sample based on the masked sample. In our experiment, we use the Data2vec Large ll60k model (D2V-L), which is pretrained on 60k LibriLight English audio.

4. Experiments

From ERJ, eight vowel-based minimal pairs of words and eleven consonant-based minimal pairs of words were selected, where 10 to 24 cross-gender speaker pairs were formed. From JRF, eighteen consonant-based minimal pairs of words were selected with 400 cross-gender speaker pairs. AE word productions and JP word productions were forced-aligned with the Montreal forced aligner [26] and the CSJ-KALDI, respectively.

All the word productions were converted to the ten kinds of speech representations, from each of which the contrastive phoneme segment, $c(v)$ or $c(w)$ in Section 2, was extracted to calculate $\overline{\mathrm{SI}}$ for each representation and each phoneme pair. Although PPG-DTW was often conducted with Bhattacharyya distance as local distance measure, which can calculate similarity between two probabilty distributions, in the experiments here, cosine distance was always used for generalization.

5. Results and discussion

Table 2 shows the averaged SI scores of AE over the cross-gender speaker pairs for each of the phoneme contrasts tested, and it also shows the minimum SI scores among the speaker pairs[2]. (a) and (b) correspond to vowel pairs, and (c) and (d) correspond to consonant pairs. In each row, the value in bold is the highest score. Table 3 shows the averaged SI scores and the minimum SI scores of JP, where only the average of the averaged IS and the average of the minimum SI are presented.

As expected, in (a) of Table 2, FBANK and MFCC show lower values than the other models, which are also lower than 1.0. This indicates that speaker gaps are acoustically larger than phoneme gaps. To achieve a high SI score in this situation, good mechanism of abstraction is needed to convert speech input into its representation. In the case of consonants, the SI values in the raw features are higher than 1.0 but generally lower than the other models. This is reasonable because it is well-known that speaker gaps are acoustically larger in vowels than in consonants. In both cases of vowel and consonant, the minimum SI scores of the raw features also tend to be lower than the others.

In (a) to (d) of Table 2, AE-PPG is a reference but supervised representation, and especially in consonants, it generally shows higher scores than others. In vowels, however, AE-PPG is outperformed on average by WavLM-L. JE-PPG, which was used with language mismatch, is also outperformed by some self-supervised representations such as HuBERT and WavLM.

WavLM-L and WavLM-B show better performances than others in English vowels. The former achieved the best average SI score, while the latter achieved the highest average minimal SI score. The size of training data for WavLM-B is much smaller than that of HuBERT, but WavLM-B outperforms HuBERT on both vowels and consonants. This result supports that the gated relative position bias, which was used in training WavLM to model local information, served to satisfy the two conditions of interest very effectively.

W2V-L and D2V-L show poor performances compared to HuBERT and WavLM. Among the three main training approaches of self-supervised models, W2V and D2V take the contrastive approach and the generative approach, while HuBERT and WavLM are predictive models. One possible explanation is that, by preparing pseudo-labels and predicting them in the training process, the models are trained to enhance abstraction and generalization, which results in strengthening sensitivity to phonemic contrasts.

W2V-XL achieved a relatively high average score in consonants. However, it is interesting that its performance is not stable. In (c) of Table 2, s-th and sh-s show extraordinarily high scores, but some of the other pairs show scores lower than 1.0. In (d) of Table 2, W2V-XL shows the lowest averaged score, which is even lower than FBANK and MFCC.

In the Japanese experiment in Table 3, JP-PPG achieved the highest average scores in both averaged SI and minimum SI. W2V-XL, which was trained with 53 languages, also achieved a high average score of averaged SI, but its performance is unstable. In W2V-XL, the average of minimum SI is the lowest. If we take non-Japanese models to analyze Japanese minimal pairs, we select WavLM-based models.

All the findings considered, out of the self-supervised representations tested, we can say that WavLM-based models are superior to the others. We can say otherwise that, in WavLM, high sensitivity to phoneme contrasts and insensitivity to speaker contrasts are directly encoded in the representation itself.

6. Conclusions

In this paper, a simple and new metric was proposed to directly measure sensitivity to phoneme contrasts and insensitivity to speaker contrasts, and various self-supervised representations were examined and compared even with acoustic representa-

[2]The reason of showing the minimum SI, i.e. the worst case, is that, generally speaking, language teachers are sensitive to and eager to know in which situations technologies are not working as expected.

Table 2: *Cross-gender analysis of sensitivity to AE phonemic contrasts and insensitivity to speaker contrasts*

(a) Average of vowel-pair SI over the cross-gender speaker pairs

pair	FBANK	MFCC	AE-PPG	JP-PPG	W2V-L	W2V-XL	HuBERT	WavLM-B	WavLM-L	D2V-L
ae-eh	0.678	0.661	1.320	1.075	0.921	1.376	1.442	1.731	**1.859**	0.915
ie-eh	0.745	0.737	**1.954**	1.256	0.902	1.502	1.405	1.763	1.918	1.031
ey-eh	1.001	1.099	**2.850**	2.292	1.043	2.119	2.671	2.683	2.838	2.248
ae-ah	0.754	0.874	2.448	1.324	1.066	1.037	2.268	**3.165**	2.873	1.187
ah-aa	0.659	0.585	1.366	0.808	0.904	0.924	1.146	1.457	**1.805**	1.388
iy-ih	0.827	0.853	2.627	1.808	0.904	1.004	2.627	3.412	**4.020**	1.416
uw-uh	0.936	1.035	2.959	1.583	0.938	1.048	1.801	2.833	**3.441**	1.309
ay-ow	0.791	0.883	3.173	1.624	0.896	1.670	2.554	2.825	**3.186**	1.921
average	0.799	0.841	2.337	1.471	0.947	1.335	1.989	2.484	**2.743**	1.427

(b) Minimum of vowel-pair SI among the cross-gender speaker pairs

pair	FBANK	MFCC	AE-PPG	JP-PPG	W2V-L	W2V-XL	HuBERT	WavLM-B	WavLM-L	D2V-L
ae-eh	0.516	0.410	**0.976**	0.765	0.715	0.215	0.624	0.903	0.649	0.492
ie-eh	0.545	0.437	**1.354**	0.724	0.692	0.191	0.545	1.059	0.723	0.510
ey-eh	0.820	0.752	1.001	0.729	0.836	0.875	0.748	**1.263**	1.201	0.872
ae-ah	0.515	0.496	1.661	0.823	0.891	0.421	1.121	**2.152**	1.742	0.861
ah-aa	0.457	0.439	0.519	0.434	**0.698**	0.321	0.304	0.615	0.606	0.685
iy-ih	0.564	0.634	1.296	0.719	0.735	0.388	0.822	1.673	**1.917**	1.119
uw-uh	0.712	0.749	1.611	1.198	0.699	0.486	1.270	**1.896**	1.885	0.641
ay-ow	0.610	0.590	1.897	1.102	0.768	0.964	1.166	**2.239**	1.860	1.060
average	0.592	0.563	1.289	0.812	0.754	0.483	0.825	**1.475**	1.323	0.780

(c) Average of consonant-pair SI over the cross-gender speaker pairs

pair	FBANK	MFCC	AE-PPG	JP-PPG	W2V-L	W2V-XL	HuBERT	WavLM-B	WavLM-L	D2V-L
b-p	2.210	1.442	**3.766**	1.628	1.120	0.843	2.423	2.718	3.249	1.777
d-t	2.019	1.390	**3.653**	1.676	0.969	0.950	2.049	2.430	2.564	2.037
g-k	1.707	1.301	**3.430**	1.921	1.125	0.908	1.581	2.111	2.132	1.529
s-th	1.130	0.960	1.808	1.208	0.996	**10.245**	1.612	1.489	1.655	1.201
dh-z	1.163	0.997	**1.744**	1.352	0.987	0.918	1.345	1.437	1.511	1.046
sh-s	0.925	1.025	3.360	1.394	0.955	**7.160**	1.628	1.868	1.803	1.272
f-h	1.867	1.707	**4.537**	3.138	1.266	1.753	2.508	3.279	4.140	2.350
b-v	1.070	0.990	1.666	1.149	1.092	1.109	1.388	**1.965**	1.900	1.390
ch-jh	1.003	0.932	**1.713**	0.953	0.940	1.218	1.148	1.220	1.106	1.115
m-n	0.646	0.624	**1.486**	0.656	0.955	0.628	0.830	1.319	1.422	0.775
l-r	0.710	0.782	1.805	0.784	0.944	1.015	1.528	2.004	**2.094**	1.341
v-w	1.202	1.260	**3.189**	1.988	1.084	0.986	1.605	1.853	1.908	1.296
average	1.304	1.118	**2.680**	1.487	1.036	2.311	1.637	1.975	2.124	1.427

(d) Minimum of consonant-pair SI over the cross-gender speaker pairs

pair	FBANK	MFCC	AE-PPG	JP-PPG	W2V-L	W2V-XL	HuBERT	WavLM-B	WavLM-L	D2V-L
b-p	0.933	1.027	**1.605**	0.997	0.770	0.118	1.044	1.357	0.887	1.094
d-t	1.173	0.871	**1.755**	0.999	0.738	0.184	0.652	1.379	0.909	1.127
g-k	1.341	0.978	**1.861**	1.272	0.894	0.253	0.685	1.307	1.217	0.993
s-th	0.798	0.777	**1.147**	0.895	0.773	0.606	0.999	0.899	0.994	0.959
dh-z	0.525	0.490	0.536	0.583	**0.667**	0.034	0.380	0.474	0.359	0.402
sh-s	0.745	0.760	**1.445**	0.554	0.785	0.031	0.528	1.054	0.523	0.638
f-h	0.963	1.039	**2.177**	1.496	0.704	0.740	1.106	1.057	1.050	0.881
b-v	0.407	0.521	0.482	0.359	0.747	0.437	0.420	0.621	**0.639**	0.618
ch-jh	0.690	0.660	1.001	0.688	0.687	0.901	0.799	**0.967**	0.710	0.607
m-n	0.470	0.452	0.638	0.266	0.723	0.134	0.371	**0.856**	0.743	0.477
l-r	0.410	0.553	**0.903**	0.498	0.584	0.651	0.530	0.810	0.850	0.581
v-w	0.860	0.891	**1.836**	1.027	0.897	0.067	0.973	0.934	0.977	0.743
average	0.776	0.751	**1.282**	0.803	0.747	0.346	0.707	0.976	0.822	0.760

Table 3: *Cross-gender analysis of sensitivity to JP consonant contrasts and insensitivity to speaker contrasts*

	FBANK	MFCC	AE-PPG	JP-PPG	W2V-L	W2V-XL	HuBERT	WavLM-B	WavLM-L	D2V-L
average of averaged SI	1.166	0.931	1.356	**2.205**	1.033	2.085	1.202	1.663	1.653	1.219
average of minimum SI	0.500	0.507	0.493	**0.760**	0.431	0.098	0.416	0.621	0.497	0.459

tions and phonetic posteriorgrams. Since we're interested in (in)sensitivity directly encoded in the representations, our metric is based on comparison of a minimal pair of words without any additional training. This metric is very useful especially to select good models for L2 speech assessment. Experiments showed that WavLM-based representations are superior to the other self-supervised representations in the task of comparison-based L2 speech assessment, and even better in some cases than phonetic posteriorgram. Although WavLM was trained only with English speech samples, it showed a good performance in Japanese experiments. If higher language independence is required, WavLM-based models trained with multiple languages may provide a better solution.

7. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, pp. 832–844, 2009.

[2] T. Kawahara and N. Minematsu, "Computer-Assisted Language Learning (CALL) based on speech technologies," *IEICE Trans. Info. Sys.*, vol. J96-D, no. 7, pp. 1549–1565, 2013.

[3] T. Isaacs, "Fully automated speaking assessments: changes to proficiency testing and the role of pronunciation," in *The Routledge handbook of contemporary English pronunciation*, O. Kang, R. I. Thomson, and J. Murphy, Eds. Routledge, 2018, pp. 570–584.

[4] Y. Yamashita, K. Kato, and K. Nozawa, "Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 496–501, 03 2005.

[5] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Proc. Spoken Language Technology*, 2012, pp. 382–387.

[6] R. Rasipuram, M. Cernak, A. Nachen, and M. Magimai-Doss, "Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities," in *Proc. INTERSPEECH*, 2015, pp. 648–652.

[7] J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, "Automatic scoring of shadowing speech based on DNN posteriors and their DTW," in *Proc. INTERSPEECH*, 2017, pp. 1422–1426.

[8] C. Zhu, N. Minematsu, and N. Nakanishi, "Multi-granularity annotation of instantaneous intelligibility of learners' utterances based on shadowing techniques," in *Proc. Automatic Speech Recognition and Understanding*, 2021.

[9] A. Sini, A. Perquin, D. Lolive, and A. Delhay, "Phone-Level pronunciation scoring for L1 using weighted-dynamic time warping," in *Proc. Spoken Language Technology*, 2023, pp. 1081–1087.

[10] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[11] Y. Kashiwagi, C. Zhang, D. Saito, and N. Minematsu, "Divergence estimation based on deep neural networks and its use for language identification," in *Proc. ICASSP*, 2016, pp. 5435–5439.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. B. Glembek, N. Goel, M. Hannemann, P. Motlíček, Q. Y., S. P., J. Silovský, G. Stemmer, and K. Veselý, "The KALDI speech recognition toolkit," in *Proc. ASRU*, 2011.

[13] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maalœ, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[14] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[15] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities," in *Proc. ACL*, 2022, pp. 8479–8492.

[16] T.-h. Feng, A. Dong, C.-F. Yeh, S.-w. Yang, T.-Q. Lin, J. Shi, K.-W. Chang, Z. Huang, H. Wu, X. Chang, S. Watanabe, A. Mohamed, S.-W. Li, and H.-y. Lee, "Superb@SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *Proc. Spoken Language Technology*, 2023, pp. 1096–1103.

[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] K. Nishina, Y. Yoshimura, I. Saita, Y. Takai, K. Maekawa, N. Minematsu, S. Nakagawa, S. Makino, and M. Dantsuji, "Development of Japanese speech database read by non-native speakers for constructing CALL System," in *Proc. ICA*, 2004, pp. 561–564.

[19] "Wall street journal corpus," https://catalog.ldc.upenn.edu/LDC93s6a, accessed: 2023-05-15.

[20] "Corpus of spontaneous japanese," https://clrd.ninjal.ac.jp/csj/en/, accessed: 2023-05-15.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[22] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-Lingual representation learning for speech recognition," in *Proc. INTERSPEECH*, 2021, pp. 2426–2430.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[25] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. ICML*, 2022, pp. 1298–1312.

[26] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using KALDI," in *Proc. INTERSPEECH*, 2017, pp. 498–502.