



# OMPAL: Bridging Speech and Learning with an Open-Source Mandarin Pronunciation Assessment Corpus for Global Learners

Wen-Wei Hsieh<sup>1</sup>, Hao-Wei Chi<sup>1</sup>, Kuan-Chen Wang<sup>1</sup>, Ping-Cheng Yeh<sup>1</sup>, Te-hsin Liu<sup>2</sup>, Chen-Yu Chiang<sup>3</sup>

<sup>1</sup>Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

<sup>2</sup>Graduate Program of Teaching Chinese as a Second Language, National Taiwan University, Taiwan

<sup>3</sup>Department of Communication Engineering, National Taipei University, Taiwan

{r11942078, pcyeh, tehsinliu}@ntu.edu.tw, cychiang@mail.ntpu.edu.tw

## Abstract

This paper introduces OMPAL, a new open-source Mandarin corpus specifically designed for non-native pronunciation assessment. This corpus comprises 1,768 Mandarin utterances from French L1 speakers learning Mandarin, each meticulously annotated by four experts with professional Mandarin teaching experience at both the word and sentence levels. We also provide a manual scoring system to assist researchers in constructing related corpora. Furthermore, a baseline model for pronunciation assessment, which is publicly accessible, is provided alongside our corpus. The OMPAL corpus, available for commercial and non-commercial use, is designed to support and enhance speech research across various applications. We believe that OMPAL will be a valuable resource for the speech research community.

**Index Terms:** corpus, Mandarin, second language (L2), deep learning, computer-aided pronunciation training (CAPT)

## 1. Introduction

Computer-assisted language learning (CALL) has seen substantial advancements in recent years, allowing learners to study foreign languages conveniently [1] and prepare for international language proficiency tests [2] through mobile applications. A critical component of CALL, Computer-aided pronunciation training technology (CAPT), fulfills two key roles: pronunciation assessment and pronunciation teaching. This technology not only corrects the mistakes of the learners, but also guides them toward producing more precise and native-like utterances [3].

Mandarin Chinese presents unique challenges for second-language learners, mainly due to the perception and production of lexical tones [4]—where pitch contours change word meanings—and the variety of retroflex, palatal, affricate, and fricative consonants with subtle differences [5, 6]. These features complicate both pronunciation and comprehension. To equip learners, we organize practical pronunciation assessment tasks into two levels: word-level, focusing on individual Chinese characters, and sentence-level, evaluating entire sentences. At the sentence level, evaluations assess overall sentence integrity using three criteria: accuracy, fluency, and prosody. Accuracy measures the correctness of each word in the sentence, fluency examines the smoothness of speech and the absence of unnecessary pauses, and prosody assesses the naturalness of intonation and cadence. The word-level assessment focuses on individual Chinese characters, evaluating consonants, vowels, and tones to determine their correctness.

Obtaining corpora for pronunciation assessment is often costly due to the extensive effort and resources required for data annotation and processing. While there are numerous sec-

ond language corpora focusing on English, only three such corpora exist for Mandarin. Among these, just one is publicly accessible, and it features recordings from merely four speakers, lacking a comprehensive assessment methodology. Given the global surge in interest in learning Mandarin, establishing publicly accessible Mandarin second language corpora has become an urgent necessity. We compared the non-native speech corpora in Table 1. Despite this comparison, there remains a significant need to increase the availability of Mandarin corpora, as most datasets stay inaccessible to the public. Moreover, the few available corpora often need more transparency in the annotation process, raising concerns about the rigor and reliability of the data for comprehensive pronunciation assessment.

Our research makes significant contributions by establishing a rigorously evaluated corpus with expert grading procedures, providing an open-source dataset, and developing a baseline model on the OMPAL corpus for future reference. These initiatives are crucial as they address the critical shortage of non-native Mandarin resources and enhance tools available for pronunciation assessment, fostering further academic research and practical applications in this field.

Below is the outline for the rest of this paper: Section 2 details the design of the corpus and its scoring system. Section 3 describes the experimental setup, including baseline models and their results. The paper concludes with Section 4, which discusses the findings and outlines future research directions.

## 2. Design of Corpus

In this section, we describe the OMPAL corpus in detail, which comprises 1,768 Mandarin utterances recorded by 46 French-speaking learners with beginner and intermediate Mandarin levels. Experts with extensive teaching experience in Mandarin have carefully annotated these recordings.

### 2.1. Scripts and Speakers

We selected the scripts for the OMPAL corpus to match the learners' proficiency levels, drawing from Chinese language textbooks provided by INALCO (Institut National des Langues et Civilisations Orientales [National Institute of Oriental Languages and Civilizations]). This selection criterion ensures the content is relevant and familiar to learners, allowing them to concentrate on refining their pronunciation and intonation. As a result, learners can produce speech more naturally. The corpus primarily features frequently used dialogues that emphasize intonation patterns and varied grammatical structures to capture differences effectively.

The corpus features recordings from 46 French-speaking Mandarin learners, representing a diverse range of proficiency levels. The participant demographics include 21 first-year stu-

Table 1: Comparison of non-native speech corpora.

Corpus	Target Languages (L2)	Native Languages (L1)	Dur(hr)#Utt	#Speakers	Open-source
ISLE [7]	English	German and Italian	18/-	46	Yes
ERJ [8]	English	Japanese	-/68,000	200	Yes
iCALL [9]	Mandarin	24 languages	142/90,841	305	No
SingaKids-Mandarin [10]	Mandarin	Singaporean (English)	125/79,843	255	No
L2-ARCTIC [11]	English	5 languages	3.6/-	24	Yes
VoisTUTOR [12]	English	6 languages	14/26,529	16	No
EpaDB [13]	English	Spanish	-/3,200	50	No
SELL-CORPUS [14]	English	Chinese	31.6/-	389	Yes
Speechocean762 [15]	English	Chinese	-/5,000	250	Yes
LATIC [16]	Mandarin	Russian, Korean, French, and Arabic	4/2,579	4	Yes
Arabic-CAPT [17]	Arabic	From 20 countries	2.3/1,611	62	No
AraVoiceL2 [18]	Arabic	From 5 countries	5.5/7,062	11	No
<b>OMPAL (proposed)</b>	<b>Mandarin</b>	<b>French</b>	<b>2.9/1,768</b>	<b>46</b>	<b>Yes</b>

dents (5 males and 16 females) and 25 second-year students (8 males and 17 females). Each participant read prescribed texts and recorded their speech using electronic devices in acoustically controlled environments. We then segmented these recordings into discrete utterances, subsequently annotated by four experts with backgrounds in Mandarin language education.

## 2.2. Expert Annotations

Expert annotations are crucial to ensuring the quality of our corpus’ pronunciation assessment. Our panel of experts consists of four seasoned instructors who specialize in teaching Chinese as a second language and are paid by the hour. They independently assessed Mandarin utterances using uniform criteria to maintain consistency and objectivity. Each utterance is evaluated by three out of four experts and can be reviewed multiple times.

The scoring metrics, developed by the experts and authors, are detailed in Table 2. At the word level, scores are assigned based on the accuracy of consonants, vowels, and tones for each word within an utterance. Experts determine the correctness of each word using these three parameters. At the sentence level, the experts assess the overall accuracy, fluency, and prosody of entire sentences. Experts have the discretion to choose the order in which they evaluate these aspects.

Second-language learners often repeat or backtrack on a word or phrase, but only the final attempt is valid. If a learner omits a word, substituting it with sounds like ”uh” due to unfamiliar pronunciation, the experts mark the consonants, vowels, and tones as incorrect (check 0). If the omission is due to forgetfulness yet other word pronunciations are over 90% correct, the experts mark it as correct (do not check); otherwise, they mark it as incorrect (check 0). For recordings with repeated words, the experts evaluate the last word at the word level and the entire utterance at the sentence level.

To facilitate efficient scoring, we developed an open-source system, as shown in Figure 1, allowing experts to evaluate utterances conveniently. Since they typically spend about an hour evaluating utterances for a time, they often become fatigued by the challenges of non-native pronunciation. To alleviate this issue, we have designed each section to include four utterances that express the same sentences in a random order. One of the utterances is provided by a native speaker to serve as an anchor for calibrating scoring standards and enhancing reliability.

Figure 2 illustrates the distribution of word-level scores, while Figure 3 displays the sentence-level score distribution. Although Mandarin learners achieve high correctness rates

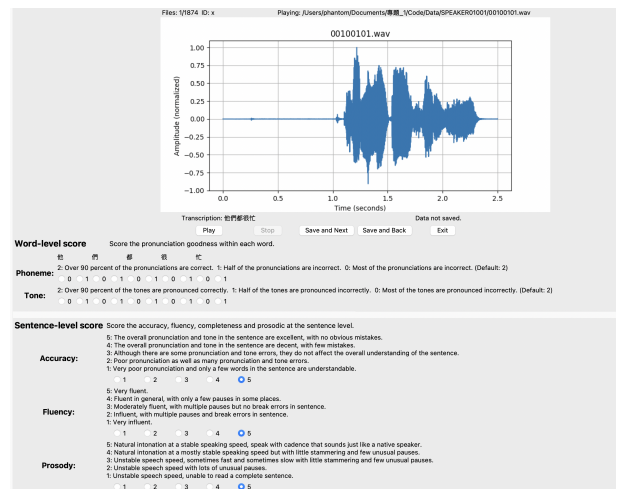


Figure 1: The open-source scoring system. Each word is evaluated on consonants, vowels, and tone in the word-level scoring process.

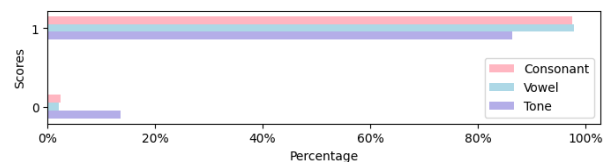


Figure 2: Word-level score distribution: 1 denotes correct pronunciation, and 0 denotes incorrect pronunciation.

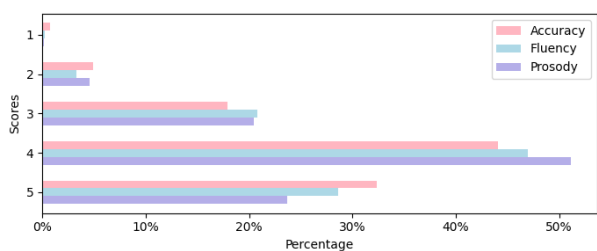
above 90% for consonants and vowels, with some tone mistakes, their performance at the sentence level—assessed in accuracy, fluency, and prosody—does not reach similarly high levels. This discrepancy highlights that native-like fluency requires consistent, accurate pronunciation at the word level and mastery of suprasegmental features such as intonation and cadence.

## 3. Experiment

This section details the experimental design used to establish baselines for the OMPAL corpus in pronunciation assessment.

Table 2: *Scoring metrics.*

Score	Description
<b>Word-level Tone/Consonant/Vowel</b>	
1	The pronunciation is correct.
0	The pronunciation is incorrect.
<b>Sentence-level Accuracy</b>	
5	The overall pronunciation and tone in the sentence are excellent, with no obvious mistakes.
4	The overall pronunciation and tone in the sentence are decent, with few mistakes.
3	Although there are some pronunciation and tone errors, they do not affect the overall understanding of the sentence.
2	Poor pronunciation as well as many pronunciation and tone errors.
1	Very poor pronunciation and only a few words in the sentence are understandable.
<b>Sentence-level Fluency</b>	
5	Very fluent.
4	Fluent in general, with only a few pauses in some places.
3	Moderately fluent, with multiple pauses but no break errors in a sentence.
2	Influent, with multiple pauses and break errors in a sentence.
1	Very inefficient.
<b>Sentence-level Prosody</b>	
5	Natural intonation at a stable speaking speed, speak with a cadence that sounds just like a native speaker.
4	Natural intonation at a mostly stable speaking speed but with little stammering and few unusual pauses.
3	Unstable speech speed, sometimes fast and sometimes slow with little stammering and few unusual pauses.
2	Unstable speech speed with lots of unusual pauses.
1	Unstable speech speed, unable to read a complete sentence.

Figure 3: *Sentence-level score distribution.*

Initially, we calculate each speaker’s average scores of all utterances precisely to segment our testing sets. This corpus includes 46 non-native speakers divided into four proficiency levels based on these average scores using quartile divisions.

To form the test sets, we select combinations of speakers, ensuring each set contains speakers from each proficiency level and includes one native speaker chosen from a pool of three. Our protocol ensures all possible combinations are considered, with each testing set comprising two males and two females. Additionally, each set must include at least two beginners and two intermediate-level speakers as part of our predefined proficiency criteria. Combinations that do not meet these gender and proficiency requirements are discarded.

We randomly select five from the remaining valid combinations to serve as the testing sets and show the first combination in Table 3. These sets are crafted to balance gender and proficiency levels, ensuring a comprehensive evaluation of our model’s performance. The speakers not included in the testing set are grouped into the training set, comprising 42 speakers. This structured approach allows for a robust assessment across varied linguistic backgrounds and proficiency levels.

Table 3: *Detailed information for Combination 1 used in the test set of the OMPAL corpus. Score is the average of accuracy, fluency, and prosody, and its maximum is 15.*

SpeakerID	Gender	Score	Proficiency level
01002	Male	15.00	Native
02021	Male	8.50	Beginner
02007	Female	10.35	Beginner
02036	Female	11.16	Intermediate
02031	Female	12.23	Intermediate

### 3.1. Training and test set

To ensure a robust evaluation of the performance of our model in the OMPAL corpus in the evaluation of pronunciation, we employed a rigorously designed experimental setup. We randomly selected five speaker combinations as test sets using a protocol that balanced gender and proficiency levels. The 46 non-native speakers of the corpus are categorized into four proficiency levels determined by quartile divisions of their average scores. For each test set combination, we chose four non-native speakers from each proficiency level and one native speaker from a pool of three available natives. Each combination consisted of a minimum of two male and two female speakers, along with at least two predefined beginners and two predefined intermediate-level speakers. Table 3 provides an example configuration of one such test set combination. The training set consisted of the remaining speakers not included in the test set, including 42 non-native speakers.

### 3.2. The structure of the proposed method

The overview of our method is shown in Figure 4. Our model utilizes the pre-trained model to convert raw audio into detailed and encoded acoustic representations. We used the pypinyin

library in Python to convert Chinese text into tone-marked pinyin strings, thereby obtaining phonetic information. Then, we mapped each character and tone to a unique integer index for numerical processing in models. The combined data streams are inputted into a neural network to estimate sentence-level scores for accuracy, fluency, and prosody. This approach effectively combines linguistic and acoustic data, providing subtle insights into speech characteristics.

Wav2vec 2.0 [19], an advanced self-supervised model, encodes speech to extract meaningful acoustic features [20]. The model comprises a convolutional encoder for converting raw audio into latent representations, a transformer-based context encoder for generating contextual embeddings, and a quantization module that optimizes contrastive loss during training. For effective feature extraction, audio files are resampled to 16 kHz and processed through a pre-trained encoder.

We extract features from each layer of Wav2Vec 2.0, as every layer contributes distinct acoustic or linguistic insights that collectively enhance speech analysis. In recent studies [21, 22], it was shown that the lower layers capture detailed acoustic and phonetic information, whereas the higher layers represent broader linguistic and contextual cues. Building on these insights, we partition the hidden states of the model into the first (layers 0–5), the middle (layers 6–17), and the last (layers 18–24) groups. For each group, we first compute the mean across the time dimension for every layer’s output, then average across layers to obtain a single representation. This multi-step aggregation preserves fine-grained acoustic details while also integrating higher-level linguistic context, thereby ensuring better performance even in data-scarce environments.

We formulated the problem as a regression task to accurately forecast pronunciation scores, minimizing L2 loss between predictions and expert labels. We utilized a Bidirectional Long Short-Term Memory (BLSTM) model to integrate acoustic and linguistic features for dynamic sentence-level scoring. Global Average Pooling was then applied to aggregate the audio and linguistic context vectors over their respective dimensions—time for audio and character sequence for text—before processing through a linear layer. Recognizing the strong correlation between audio duration and performance, we incorporated duration measurements and concatenated this data with other features in the final linear layer to enhance prediction.

### 3.3. Experimental Setup

Our experiments used the pre-trained 'wav2vec2-large-960h' model from HuggingFace for feature extraction. We implemented a double-layer BLSTM network with 256 hidden units and 128 embedding dimensions [23, 24]. The model optimization was performed on the BLSTM using the Adam optimizer with the L2 Loss function, and it was trained over 40 epochs with a batch size of 4 on an Nvidia GeForce RTX 4090 GPU. The learning rate followed a Reduce-on-Plateau schedule: if validation performance failed to improve for five epochs, the rate was reduced to 90% of its previous value. We evaluated the system with five-fold speaker-stratified cross-validation, reporting Pearson’s Correlation Coefficient (PCC) and Mean-Squared Error (MSE) between the predicted scores and expert annotations.

### 3.4. Results

We assessed our model by comparing its predictions to expert annotations. Table 4 summarizes our findings: even with just 1,768 utterances, our proposed model delivers decent correla-

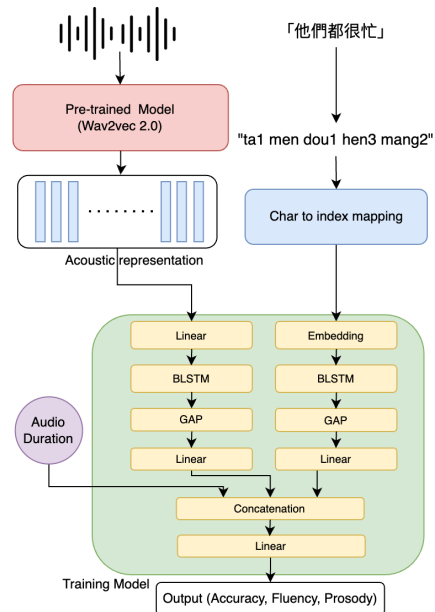


Figure 4: Overall procedure of the proposed method.

Table 4: Performance metrics across different combinations through 5-fold cross-validation, presented as mean values.

Comb. \ Metrics	Accuracy		Fluency		Prosody	
	MSE	PCC	MSE	PCC	MSE	PCC
<b>Comb. 1</b>	0.33	0.57	0.24	0.71	0.30	0.62
<b>Comb. 2</b>	0.37	0.54	0.22	0.74	0.26	0.60
<b>Comb. 3</b>	0.50	0.59	0.12	0.82	0.21	0.73
<b>Comb. 4</b>	0.72	0.64	0.27	0.82	0.33	0.79
<b>Comb. 5</b>	0.53	0.19	0.28	0.55	0.32	0.42
<b>Overall Avg.</b>	0.49	0.51	0.23	0.73	0.28	0.63

tions and error in modeling non-native Mandarin pronunciation.

Three factors likely held performance below its full potential. The wav2vec2 backbone was trained on English data, which can dull its sensitivity to Mandarin’s tonal and segmental nuances. Our training set was also unbalanced across speakers, while the test folds were perfectly stratified—introducing a distribution shift that can undermine generalization. Finally, the modest size of our corpus limits exposure to the full spectrum of learner accents and error patterns. In future work, Mandarin-focused feature extractors, more balanced sampling, and expanded data collection should help drive MSE even lower and PCC even closer to 1.

## 4. Conclusion

In this paper, we present significant advancements within CAPT. By developing and evaluating the OMPAL corpus, we have provided a valuable resource for non-native Mandarin pronunciation assessment. Our approach leads to a baseline system capable of accurately predicting pronunciation scores. These efforts address the scarcity of resources in non-native Mandarin learning and set a new standard for pronunciation assessment tools, promising to significantly impact both academic research and practical applications in language learning technologies.

## 5. References

- [1] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "Eduspeak@: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.
- [2] L. Gu, L. Davis, J. Tao, and K. Zechner, "Using spoken language technology for generating feedback to prepare for the toefl ibt@ test: A user perception study," *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 1, pp. 58–76, 2021.
- [3] Y. El Kheir, A. Ali, and S. A. Chowdhury, "Automatic pronunciation assessment-a review," *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [4] N. F. Chen, V. Shivakumar, M. Harikumar, B. Ma, and H. Li, "Large-scale characterization of mandarin pronunciation errors made by native speakers of european languages." in *Interspeech*, 2013, pp. 2370–2374.
- [5] Y.-h. Lai, "Asymmetry in mandarin affricate perception by learners of mandarin chinese," *Language and cognitive processes*, vol. 24, no. 7-8, pp. 1265–1285, 2009.
- [6] X. Wang and J. Chen, "The acquisition of mandarin consonants by english learners: The relationship between perception and production," *Languages*, vol. 5, no. 2, p. 20, 2020.
- [7] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The isle corpus of non-native spoken english," in *Proceedings of LREC 2000: Language Resources and Evaluation Conference*, vol. 2. European Language Resources Association, 2000, pp. 957–964.
- [8] N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, "Measurement of objective intelligibility of japanese accented english using ERJ (english read by japanese) database." in *INTERSPEECH*, 2011, pp. 1481–1484.
- [9] N. F. Chen, R. Tong, D. Wee, P. X. Lee, B. Ma, and H. Li, "iCALL corpus: Mandarin chinese spoken by non-native speakers of european descent." in *INTERSPEECH*, 2015, pp. 324–328.
- [10] G. Shang and S. Zhao, "Singapore mandarin: Its positioning, internal structure and corpus planning," in *Paper presented at the 22nd Annual Conference of the Southeast Asian Linguistics Society, Agay, France*, 2012.
- [11] G. Zhao, E. Chukharev-Hudilainen, S. Sonaat, A. Silpachai, I. Lucic, R. Gutierrez-Osuna, and J. Levis, "L2-ARCTIC: A non-native English speech corpus," *Proc. Interspeech 2018*, pp. 2783–2787, 2018.
- [12] C. Yarra, A. Srinivasan, C. Srinivasa, R. Aggarwal, and P. K. Ghosh, "VoisTUTOR corpus: A speech corpus of indian l2 english learners for pronunciation assessment," in *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2019, pp. 1–6.
- [13] J. Vidal, L. Ferrer, and L. Brambilla, "Epadb: A database for development of pronunciation assessment systems." in *INTERSPEECH*, 2019, pp. 589–593.
- [14] Y. Chen, J. Hu, and X. Zhang, "SELL-CORPUS: an open source multiple accented chinese-english speech corpus for l2 english learning assessment," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7425–7429.
- [15] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "Speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.
- [16] X. ZHANG, "LATIC: A non-native pre-labelled mandarin chinese validation corpus for automatic speech scoring and evaluation task," 2021.
- [17] M. Algabri, H. Mathkour, M. Alsulaiman, and M. A. Bencherif, "Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native arabic speech," *Mathematics*, vol. 10, no. 15, p. 2727, 2022.
- [18] Y. E. Kheir, S. A. Chowdhury, A. Ali, H. Mubarak, and S. Afzal, "Speechblender: Speech augmentation framework for mispronunciation data generation," *arXiv preprint arXiv:2211.00923*, 2022.
- [19] 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.
- [20] S.-w. 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 *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.
- [21] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.
- [22] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, p. 3400–3404, 2021.
- [23] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 338–345.
- [24] L. Chen, J. Tao, S. Ghaffarzagdegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6234–6238.