
Proceedings Paper:
Islam, E. orcid.org/0000-0002-5329-0414, Park, C. orcid.org/0000-0001-6671-1671 and Hain, T. (2023) Exploring speech representations for proficiency assessment in language learning. In: 9th Workshop on Speech and Language Technology in Education (SLaTE) Proceedings. 9th Workshop on Speech and Language Technology in Education (SLaTE), 18-20 Aug 2023, Dublin, Ireland. International Speech Communication Association (ISCA) , pp. 151-155.

https://doi.org/10.21437/slate.2023-29

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

Exploring Speech Representations for Proficiency Assessment in Language Learning

Elaf Islam, Chanho Park*, Thomas Hain*

Dept. of Computer Science, The University of Sheffield, Sheffield, United Kingdom
{ejislam1,cpark12,t.hain}@sheffield.ac.uk

Abstract

Automatic proficiency assessment can be a useful tool in language learning, for self-evaluation of language skills and to enable educators to tailor instruction effectively. Often assessment methods use categorisation approaches. In this paper an exemplar based approach is chosen, and comparisons between utterances are made using different speech encodings. Such an approach has the advantage to avoid formal categorisation of errors by experts. Aside from a standard spectral representation pretrained model embeddings are investigated for the usefulness for this task. Experiments are conducted using speechocean762 database, which provides 3 levels of proficiency. Data was clustered and performance of different representations is assessed in terms of cluster purity as well as categorisation correctness. Cosine distance with Whisper representations yielded better clustering performance.

Index Terms: self-supervised representation, proficiency assessment, language learning

1. Introduction

Achieving proficiency in a language involves developing skills in various linguistic domains, such as vocabulary, grammar, listening comprehension, reading comprehension, writing proficiency, and spoken language. Language learning requires the assessment of proficiency. Since recruiting and training new human experts is expensive and yields only a small performance enhancement, automatic graders can provide greater consistency and speed at a lower cost [1]. The input sequence data from a student is utilised in automated assessment of L2 spoken language proficiency. This allows for the prediction of level with regard to the student's overall proficiency as well as particular parts of their ability [2].

Automatic Speech Recognition (ASR) uses training on speech data and incorporates phonetic dictionaries to accurately recognize and transcribe phonemes, phonetics, pitch, intonation, duration, and pronunciation for assessment purposes. However, achieving good performance with ASR requires a substantial amount of transcribed audio data [3]. An ASR system employs the language model to overcome faulty acoustics to produce the correct letter sequence. Another model for language assessment is a mispronunciation detection and diagnosis (MDD) model. Language model restrictions will overlook mispronunciations. Strong acoustic modelling is needed to distinguish native products with canonical phonetic pronunciations from non-native pronunciations [4],[5]. Goodness of Pronunciation (GOP) is another popular pronunciation measure. Witt et al. [6] used Gaussian mixture model-hidden Markov model (GMM-HMM) based native acoustic model to define GOP and compute a score from the formulated GOP. Following that, the majority of the works improved either by proposing variants to the GOP-based formulation or by improving the quality of the native acoustic models [7]. In [8], advanced second language speakers are automatically oral proficiency tested. Students read, repeat, and record using a spoken dialogue system. Human ratings of their speech proficiency are compared to automatic indicators. Unlike other research, posterior scores do not correlate with human reading exercise assessments. All previous approaches either extracted sets of hand-crafted features related to specific aspects of proficiency, such as fluency, pronunciation, and prosody, or concatenated multiple features targeting multiple aspects, which were then fed into graders to predict analytic scores targeting those specific aspects.

Self-supervised learning (SSL) has recently shown promising results in speech processing applications [9, 10, 11, 12, 13, 14]. SSL can learn rich speech representations without transcription labels by training on massive unlabeled audio data. This method is better at handling different speech patterns and conditions because it captures many acoustic and linguistic features. Some previous research has explored the use of SSL in the domain of proficiency assessment. In [15], the feasibility of using Wav2vec 2.0 representations to assess L2 spoken English proficiency holistically and analytically with limited data is investigated. The study demonstrates that the Wav2vec 2.0 approach surpasses the BERT baseline system [16] in classifying Common European Framework of Reference (CEFR) levels [17]. Additionally, it shows significant improvements when utilising Wav2vec 2.0 for regression tasks targeting holistic scores in the B1 section of TLT-school, outperforming the BERT baseline trained on ASR and manual transcriptions. Another related work [18] extends a novel proficiency assessment approach using a Wav2vec 2.0 based grader on a large L2 learner dataset. The proposed approach showed good performance on parts with short spontaneous answers but faced challenges in assessing higher levels of language proficiency, as defined by the proficiency scales of the CEFR for languages [19]. In the CEFR, each speaker is graded on a scale from 1 to 6.

This paper aims to explore novel approaches for assessing spoken language proficiency using advanced speech representations. In Section 2, we introduce different speech representations, including Mel frequency cepstral coefficient, Wav2vec 2.0, and Whisper. Section 3 is dedicated to distance measurement. Moving on, Section 4 describes the data used, covering proficiency calculation, sample selection, and the features extraction. The experimental setup is detailed in Section 5, which includes the encoders employed, the clustering, and the evaluation metrics used. The results and findings are presented in Section 6, and the conclusions are provided in Section 7.

* Equal contribution.

10.21437/SLaTE.2023-29

2. Speech Representation

In this section, we discuss three different speech representations employed in our study: Mel frequency cepstral coefficient, Wav2vec 2.0, and Whisper. Each representation offers unique advantages and contributes to the accurate assessment of proficiency in language learning.

2.1. Mel frequency cepstral coefficient

The Mel frequency cepstral coefficient (MFCC) has proven to be a valuable speech representation technique for feature extraction in various speech processing tasks [20],[21]. It turns a raw audio waveform into a visual representation of the power or magnitude of various frequencies over time. It employs the Mel scale to approximate the response of the human auditory system to frequencies. The process involves applying the Short-Time Fourier Transform, converting frequencies to the Mel scale, dividing it into triangular filters, computing energy within each filter, and optionally applying a logarithmic transformation[22].

2.2. Wav2vec 2.0

Wav2vec 2.0 is an self-supervised speech representation model that uses a convolutional neural network (CNN) to learn powerful representations from large amounts of unlabeled speech data. Wav2Vec 2.0 has three main components: a convolutional feature encoder for extracting meaningful representations from raw audio, a transformer-based context encoder for capturing dependencies and temporal context, and quantisation and loss functions for discretising audio features. These components enable powerful feature extraction from raw waveforms for various speech and audio processing tasks [10].

2.3. Whisper

Whisper is a pre-trained ASR system proposed in [23]. The system was trained on 680k hours of weakly supervised datasets using multilingual and multitask learning, such as language identification, voice activity detection, speech recognition and translation. With a sequence-to-sequence transformer model predicting the next-token, for example, classification targets, Whisper was trained on multiple tasks simultaneously. The results on speech recognition showed that the system was not only robust on various target datasets because of the diversity of audio quality of training datasets but also comparable to human performance on transcribing.

3. Distance Measurement

Two different similarity measurements are used in this work based on two types of utterance representation. An utterance is either represented by features extracted at a fixed rate, or in the form of an overall embedding. Similarity between vector representations often uses the cosine distance whereas a common solution to compute distances between speech sounds is the use of Dynamic Time Warping (DTW), an example is Mel cepstral distortion as used for example in voice conversion assessment.

3.1. Cosine distance

High dimensional vector representations are commonly compared by ignoring any length variation as this may often be the results of randomness. Therefore, the cosine distance, as given by

$$\text{cd}(\mathbf{A}, \mathbf{B}) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2} \tag{1}$$

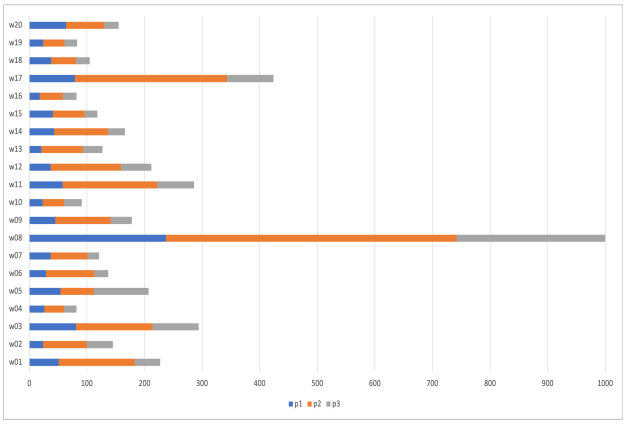


Figure (1) *Different levels of proficiency and the incidence of words in each level. p1, p2 and p3 denote low, medium and high proficiency respectively.*

is used, where $\mathbf{A}$ and $\mathbf{B}$ are vectors, $\mathbf{A}\cdot\mathbf{B}$ is a dot product of $\mathbf{A}$ and $\mathbf{B}$, and $\|\mathbf{A}\|_2$ is 2-norm of $\mathbf{A}$.

3.2. DTW or Mel cepstral distortion

DTW allows the alignment of two sequences of different length, and via that alignment, to compute a distance between such sequences. The path of alignment yielding the smallest overall distance between the sequences, represented by feature vectors, is found. The total distance represents the match. Several different types of distance functions between features can be used, including Euclidean and cosine distance, or in the case of vectors representing distributions, Kullback Leibler Divergence. The special case of Mel cepstral distortion makes use of MFCCs [24], and is defined as

$$\text{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{i=1}^{k}(m_i^c - m_i^t)^2} \tag{2}$$

where, i represents mel cepstral coefficients index, m_i^c and m_i^t denote i^{th} dimensional coefficient of the converted and target coefficients, respectively. Note that the total MCD score is length normalised to enable comparability between samples.

4. Data

In this study, the speechocean762 corpus was used, which is a free, open-source corpus of 5,000 English utterances collected from 250 Mandarin speakers [25]. The training set contains 2,500 utterances, 15,849 words, and 47,390 phones, while the test set contains 2,500 utterances, 15,967 words, and 47,688 phones. It assigns an utterance-level attribute score ranging from 0 to 10 to each speech for accuracy, fluency, completeness, prosody, and overall score. Additionally, it assigns three word-level attribute scores, including accuracy, stress, and overall score, ranging from 0 to 10 for each word. Furthermore, it provides an accuracy ranking ranging from 0 to 2 for each phoneme. Each score has been annotated by five human evaluation experts.

4.1. Proficiency calculation

The proficiency of each speaker was determined by calculating their average word accuracy across all their utterances after

combining the training and test sets. To assess proficiency in L2 learning, the study categorised the speakers into three groups: low proficiency (p1), medium proficiency (p2), and high proficiency (p3). Table 1 presents the proficiency ranges, speaker counts, and word counts for p1,p2 and p3.

Table (1) *Proficiency range, Number of speakers, Number of utterance and number of words in p1 and p2*

	Proficiency range	#Speakers	#Utterance	#Words
p1	3.9 - 7.5	52	880	7440
p2	7.6 - 8.5	69	1960	8042
p3	8.5 - 9.5	51	680	6139

4.2. Sample selection

We chose a sample of speakers aged 10 and up to ensure the quality of the speech data and to overcome potential issues with children's speech. Therefore, we extracted a diverse mix of male and female participants, totaling 176 speakers. From the adult utterances set, we selected 20 random words that had the highest occurrence rates. Figure 1 displays the different levels of proficiency and the incidence of each of the 20 words in the word list within each proficiency level.

Additionally, we ensured that there were at least ten samples of each word in each of the categories p1, p2, and p3. These selected words from the word list. Consequently, a total of 600 samples were obtained, which were further divided into two categories: reference and cluster. Specifically, for the reference, there were 120 samples available for each of the 20 words, with 2 samples representing each of the three proficiency levels. For the cluster, we utilised the same set of 20 words, with 8 samples allocated for each word and representing the three proficiency levels. Therefore, in total, we obtained 480 samples for the cluster.

4.3. Proficiency features

For clustering purposes each sample is represented by a set of features representing the relationship with the proficiency levels. Given a set of N reference samples the average utterance distance between the sample and a set of reference patterns is computed. Three feature values, one for each proficiency level p, are computed for every i-th sample of every word w, namely:

$$f^{p_i,w} = \frac{1}{N}\sum_{j=1}^{N} \text{distance}(\mathbf{S}^{p_i,w}, \mathbf{R}_j^{p_i,w}) \quad (3)$$

$\mathbf{S}^{p,w}$ and $\mathbf{R}^{p,w}$ are vectors or matrices representing utterances. N is the number of references of a word in each proficiency level and the $distance(\cdot,\cdot)$ is a distance function, as mentioned above. Thus, one sample of a word w is represented by ($f^{p_1,w}$, $f^{p_2,w}$, $f^{p_3,w}$). A slightly more complex extension to this approach might be the use of GMM-HMM-based likelihood scores. However, a nearest neighbour approach is used instead to better capture the specific nature of some error patterns.

5. Experimental setup

In this section, we provide an overview of the experimental setup employed in our study for speech representation and clustering. We utilised three different encoders: MFCC, Wav2vec 2.0 [11], and Whisper [23], to extract features from the speech data. We use the code for MCD calculation provided in the ESPnet toolkit [26]. Additionally, we employed the K-means clustering algorithm to group the samples based on their feature distances. The evaluation of the clustering was performed using cluster purity and accuracy metrics. Figure 2 illustrates the methodology employed in this project.

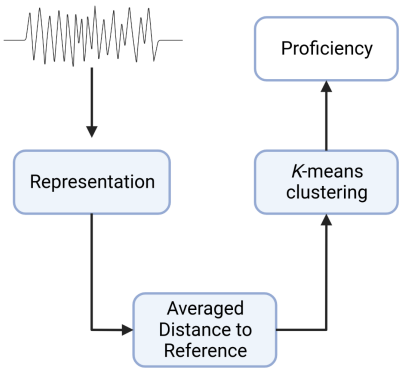


Figure (2) *An overview of the experimental setup employed in this study.*

5.1. Encoders

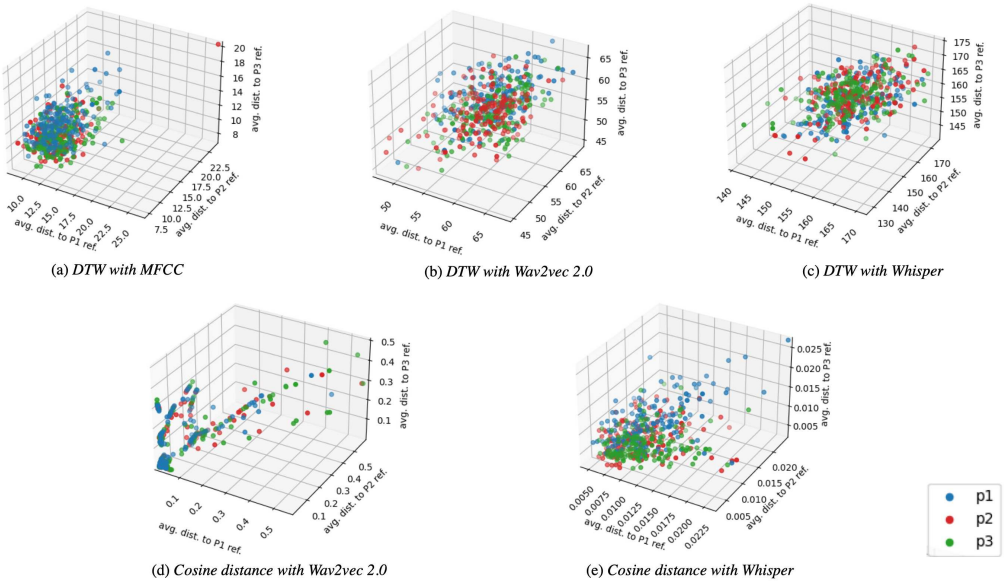
For speech representation, MFCC, Wav2vec 2.0 and Whisper were used. First, 39 features of MFCC were extracted. Maximum 100 frames among them were use to calculate the MCD after removing similar frames in sequence. Second, Wav2vec 2.0 was the model whose output is 1024 dimensions, pre-trained on Libri-Light [27], CommonVoice [28], Switchboard [29] and Fisher [30], and fine-tuned on 960 hours of LibriSpeech [31]. The utterance-level representation was obtained by averaging over frame-level outputs. Lastly, the output, 1024 dimensions, of Whisper medium system was used. The frame-level outputs were averaged to generate the utterance-level representation.

5.2. Clustering

K-means clustering implemented in [32] starts measuring the distance between centroids and data. The initial centroids can be selected either randomly or manually. According to the results of the preliminary experiment, random initial centroids of 10 were chosen for MCD and fixed ones for cosine distance. The fixed centroids were the samples whose features are close to each proficiency level, for example, (0,1,1), (1,0,1), (1,1,0). The number of clusters was set to 3, and maximum iterations were limited to 300.

5.3. Evaluation

Cluster purity and accuracy were used as evaluation metrics of the clustering. Purity is the ratio between the sum of the majority in each cluster and the number of all samples. While the cluster purity is useful to determine the number of clusters, the accuracy for clustering can be measured by the correctness rate with respect to labels assigned to the clusters if the optimal number of clusters is known.



(a) *DTW with MFCC* (b) *DTW with Wav2vec 2.0* (c) *DTW with Whisper*

(d) *Cosine distance with Wav2vec 2.0* (e) *Cosine distance with Whisper*

Figure (3) *Samples with features. p1, p2 and p3 denote low, medium and high proficiency, respectively.*

6. Results

6.1. Sample distribution in feature spaces

Samples with features average distances to references in each proficiency level are plotted in Figure 3a, 3b and 3e. While an outlier is observed at the right top corner in Figure 3a, the samples are relatively well distributed, for example, in Figure 3b. On the other hand, the samples are skewed in Figure 3d whose features are the average cosine distance of utterance-level representations. In Figure 3e, the p1 and p2 samples are spread to different directions.

6.2. Purity scores and accuracy

Purity scores and accuracy values of the clusters built by the K-means algorithm in different representation spaces are presented in Table 2. For the MFCC representation, random initial centroids were utilised, while pre-defined centroids were employed for cosine distance clustering. The cluster purity and accuracy of Wav2vec 2.0 with DTW were higher than Whisper. However, the results were opposite with cosine distance as the distribution of samples in the space of Wav2vec 2.0 with cosine distance has been heavily skewed. The samples with features using cosine distance of Whisper representation clustered better than the others although the utterance representation was averaged over frames.

Table (2) *Purity score and accuracy of clusters built using K-means in different representation spaces.*

Distance	Repr.	Init. centroids	Purity	Accuracy
DTW	MFCC	Random	0.427	0.425
	Wav2vec2		0.422	0.422
	Whisper		0.383	0.383
Cosine	Wav2vec2	Fixed	0.343	0.343
	Whisper		**0.429**	**0.429**

7. Conclusions

In conclusion, this study concentrated on automatic proficiency assessment in language learning via an exemplar-based approach and various speech representations. The goal was to avoid the need for expert categorisation of errors and to investigate alternative methods for assessing language skills.

For conducting experiments and evaluating the performance of various representations, the speechocean762 database, which provides three levels of proficiency, was used. In this study, three speech representations were used: MFCC, Wav2vec 2.0, and Whisper. As an evaluation metric, cluster purity was used to indicate the consistency of samples within each cluster. The accuracy of clustered data was also measured by comparing it to the assigned labels if the optimal number of clusters was known. Wav2vec 2.0 with DTW outperformed Whisper in terms of cluster purity and accuracy. However, for cosine distance clustering, Wav2vec 2.0 showed skewed sample distribution, resulting in suboptimal results. Notably, despite averaging utterance representations over frames, cosine distance with Whisper representation yielded better clustering performance than MFCC with DTW by 0.002 and 0.004 of purity and accuracy, respectively. The present work has certain constraints that should be addressed in future research.

Due to the limitation of the amount of data, the focus of this study was primarily on the word level, whereas language proficiency could be better measured at the sentence level. Furthermore, the examination of a limited word list might not provide a comprehensive understanding of proficiency. Therefore, expanding the word list and including words with problematic phonemes for Chinese learners would enhance the analysis. Additionally, incorporating more advanced distance measurements could facilitate more thorough analysis in future studies. Further research can build upon these findings to develop more robust and effective models for automatic proficiency assessment.

8. References

[1] M. Zhang, "Contrasting automated and human scoring of essays," *R & D Connections*, vol. 21, no. 2, pp. 1–11, 2013.

[2] N. H. De Jong, M. P. Steinel, A. F. Florijn, R. Schoonen, and J. H. Hulstijn, "Facets of speaking proficiency," *Studies in Second Language Acquisition*, vol. 34, no. 1, pp. 5–34, 2012.

[3] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[4] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[5] 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 *ICASSP 2016-2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6135–6139.

[6] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[7] S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh, "An improved goodness of pronunciation (GoP) measure for pronunciation evaluation with DNN-HMM system considering HMM transition probabilities."

[8] P. Müller, F. d. Wet, C. v. d. Walt, and T. Niesler, "Automatically assessing the oral proficiency of proficient l2 speakers," in *International Workshop on Speech and Language Technology in Education*, 2009.

[9] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[10] 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.

[11] W.-N. Hsu *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[12] Y. Meng *et al.*, "On compressing sequences for self-supervised speech models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1128–1135.

[13] Z. Chen *et al.*, "Speech separation with large-scale self-supervised learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, "Automatic pronunciation assessment using self-supervised speech representation learning," in *Proc. Interspeech 2022*, 2022, pp. 1411–1415.

[15] S. Bannò and M. Matassoni, "Proficiency assessment of L2 spoken English using wav2vec 2.0," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1088–1095.

[16] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[17] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, 2001.

[18] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. J. Gales, "L2 proficiency assessment using self-supervised speech representations," *arXiv preprint arXiv:2211.08849*, 2022.

[19] D. L. I. Persons, "Common european framework of reference for languages: Learning, teaching, assessment," 2001.

[20] A. Aggarwal *et al.*, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, 2022.

[21] C.-H. H. Yang *et al.*, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6523–6527.

[22] A. Meghanani, C. Anoop, and A. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 670–677.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[24] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." in *SLTU*, 2008, pp. 63–68.

[25] J. Zhang *et al.*, "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," *arXiv preprint arXiv:2104.01378*, 2021.

[26] S. Watanabe *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[27] J. Kahn *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[28] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[29] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[30] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text." in *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC 2004)*, vol. 4, 2004, pp. 69–71.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP 2015-2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.