



Voice-Based Dysphagia Detection: Leveraging Self-Supervised Speech Representation

Injune Hwang¹, Jung-Min Kim^{†3}, Ju Seok Ryu^{§3}, Kyogu Lee^{†1,2}

¹Department of Intelligence and Information, Seoul National University, Republic of Korea

²Artificial Intelligence Institute, Seoul National University, Republic of Korea

³Department of Rehabilitation Medicine, Seoul National University Bundang Hospital, Republic of Korea

dlswns8@snu.ac.kr, owljm@snu.ac.kr, jseok337@snu.ac.kr, kglee@snu.ac.kr

Abstract

This study introduces a framework for diagnosing dysphagia using self-supervised speech representation learning (SSL) models. Previously reported methods typically rely on mel spectrograms; however, due to the limited amount of medical data, they struggle to accurately diagnose dysphagia from low-dimensional features. However, SSL models, trained on large-scale speech data, are well suited for tasks with smaller dataset. Employing SSL features significantly enhances model performance, allowing for the model's size reduction while outperforming larger models based on mel spectrograms. Although a decrease in specificity was observed, recall, a crucial metric for disease diagnosis, showed a marked improvement, leading to a general improvement in diagnostic accuracy. Among the SSL models evaluated, the features of the 10th layer of WavLM had the highest performance. Additionally, increasing the size of the filter in the convolutional layers does not contribute to performance gains.

Index Terms: speech representation, self-supervised feature, dysphagia-aspiration detection, non-invasive diagnosis

1. Introduction

Dysphagia, characterized by impairments in the swallowing process, is often associated with neuromuscular dysfunction and strongly correlated with aging. [1, 2, 3] When dysphagia occurs, patients may experience fear of swallowing (phagophobia) or swallowing pain (odynophagia), making it difficult to maintain adequate nutritional intake. This can lead to malnutrition and complicate treatment of comorbidities. [4, 5, 6] Moreover, ineffective swallowing can result in pharyngeal residue or aspiration, potentially causing aspiration pneumonia and, in severe cases, leading to mortality. [4, 6]

The most commonly used standardized test method for dysphagia in clinical settings is the videofluoroscopic swallowing study (VFSS). [4, 7] This method involves observing the chewing and swallowing of various textured foods and liquids containing fluorescent substances such as barium through videofluoroscopy. [8, 9] Through this examination method, clinicians determine the severity of dysphagia using indicators such as the penetration-aspiration scale (PAS). The PAS indicator primarily assesses whether the ingested food touches the vocal cords or enters the airway, broadly categorizing the results as

healthy, penetration, or aspiration, using a more detailed eight-level scale. [10, 11] There are also other diagnostic methods used in research settings, such as fiberoptic endoscopic evaluation of swallowing (FEES) and manometry. [4, 12] In addition to these invasive methods, various non-invasive approaches (e.g. the gugging swallowing screen, and 3-ounce water swallow test) have been developed. [13] However, existing diagnostic methods are limited by their need for expert intervention, and specialized facilities, while also carrying risks such as radiation exposure. Owing to these limitations, periodically monitoring changes in a patient's condition during daily dietary life is challenging. [13, 14]

Considering these challenges, particularly the difficulty of periodic monitoring, several studies have explored the potential of detecting dysphagia through patients' voices. [15, 16, 17, 18, 19, 20, 21, 22, 23] Early studies on voice analysis in dysphagia patients focused on reporting statistically significant differences in parameters such as frequency perturbation (e.g. relative average perturbation (RAP)), amplitude perturbation (e.g. shimmer percentage (SHIM)), and noise-to-harmonics ratio (NHR) between healthy (low-risk group) and dysphagia (high-risk group) patients, as diagnosed by conventional dysphagia assessment methods. [15, 16] Furthermore, regarding voice changes before and after swallowing, another previous study reported that RAP showed the highest statistical significance in repeated measures mixed ANOVA statistics considering both time (pre- and post-swallowing) and group (non-aspiration risk and aspiration risk groups). [17]

In addition to statistical analyses, studies have applied classical machine learning models, such as SVM and XGBoost, to utterance-level features extracted from patients' voices, including frequency perturbation, amplitude perturbation, and harmonics-to-noise ratio (HNR). [18, 19, 20] Advancing further, there have been attempts to develop systems that can be directly applied to patients' daily lives without primary expert intervention by applying deep learning models to voice itself through frame-level features such as mel frequency cepstral coefficients (MFCCs) and mel-spectrograms, rather than using utterance-level indicators. [21, 22, 23]

As self-supervised speech representation models have advanced, there has been an increasing shift from hand-crafted features to self-supervised learning(SSL) features across various downstream speech tasks. This is particularly beneficial in scenarios where it is challenging to collect large datasets, such as speech recognition for low-resource languages or medical-data-related tasks. Leveraging information from SSL features pre-trained on large-scale speech datasets has proven its usefulness in a range of downstream tasks within the speech domain, including emotion recognition, speech enhancement, keyword spotting, and text-to-speech. [24, 25, 26, 27] To the best of our

[†] Also affiliated with Interdisciplinary Program in Artificial Intelligence, Seoul National University.

[‡] Also affiliated with Business Incubation Center, Department of Research Planning, Biomedical Research Institute, Seoul National University Bundang Hospital.

[§] Also affiliated with Seoul National University College of Medicine.

knowledge, SSL features have not yet been applied to dysphagia detection. However doing so could offer an excellent solution to address the challenges posed by limited datasets.

We propose a deep learning model for diagnosing dysphagia through voice analysis, which is a non-invasive method. Although previous models [23] have employed hand-crafted features, such as mel spectrograms, the limited amount of available data has made it challenging for these models to learn the critical patterns necessary to address complex issues such as dysphagia from low-dimensional features. In this study, we aim to achieve high performance with a small dataset by leveraging features from self-supervised speech representation models pre-trained on large-scale speech data. Our contributions include presenting a deep learning framework for diagnosing dysphagia from the voice, investigating speech features suitable for dysphagia diagnosis, and examining the contributions of pre- and post-swallowing voice features within this framework.

2. Method

2.1. Features

In this study, we compared hand-crafted features, specifically mel spectrograms, with self-supervised speech features for diagnosing dysphagia. Mel spectrogram captures the frequency characteristics of a signal over time and enhances the resolution of frequency bands that are more critical to human speech using filter banks. Mel spectrogram is used in a logarithmic scale to compress the range of values. Although mel spectrograms have been effective across various speech-related tasks and remain widely used, they are increasingly being supplanted by features derived from self-supervised learning models. The SSL models, that are pre-trained on large-scale speech data, can compensate for the scarcity of dysphagia-specific data.

Among the SSL models, we focused on three prominent examples: Wav2Vec 2.0 [28], HuBERT [29], and WavLM [30]. These models have demonstrated high performance across a range of benchmarks, suggesting their potential utility for diagnosing dysphagia as well. All three models are based on the Transformer architecture, which is well-suited for processing large-scale speech data. Despite sharing a similar foundational structure, these models employ different feature learning strategies, allowing us to compare and analyze their performances on a common architecture. Wav2Vec 2.0 learns by predicting masked portions of the speech signal, HuBERT uses the clustering of speech signals to learn hidden units, and WavLM extends the HuBERT approach by incorporating data augmentation and considering a variety of speech tasks during training. In particular, the WavLM emphasizes its non-ASR task performance.

All audio was resampled to 16 kHz before feature extraction. For the mel spectrogram, we used 128 filter banks with a window size of 640 and a hop size of 160, resulting in a 128-dimensional vector sequence at a frequency of 100 Hz. The features from Wav2Vec 2.0, HuBERT, and WavLM yielded 768-dimensional vector sequences at 50 Hz. We compared the output from all Transformer layers, from layer 0 to layer 12.

2.2. Architectures

In a baseline study [23], MobileNet-V3 [31] was used as the base model, with a final convolutional layer to reduce the number of channels for binary classification. MobileNet-V3 was chosen because it has been successfully applied to mel spectrograms in [23], demonstrating its effectiveness in this context. Designed to perform image classification efficiently with lim-

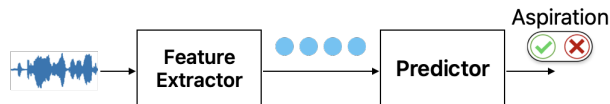


Figure 1: Overview of the voice-based dysphagia detection pipeline. The pipeline begins with a feature extractor, which can utilize various representations such as mel spectrograms, Wav2Vec 2.0, HuBERT, or WavLM. The extracted features are subsequently fed into a predictor for dysphagia classification.

ited resources, the architecture was determined through Neural Architecture Search. It incorporates inverted residual blocks, squeeze-and-excitation blocks, and a hardwired activation function to enhance both efficiency and performance. To scale the model, as in [32], the model width multiplier was set to 2.0, resulting in a parameter count of 11.65 million, thereby increasing the model’s capacity and allowing for a performance comparison with the improved model.

For models using SSL features as input, preliminary experiments indicated that high model complexity was not necessary. Unlike mel spectrograms, SSL features have a higher dimensionality, making it more challenging for complex models to learn effectively from a limited dataset. Therefore, to minimize additional complexity and leverage the rich information already captured by the SSL features, a single 1D convolutional layer was chosen. After convolution, temporal average pooling was applied to aggregate the features across the time dimension, followed by a linear layer to produce the final binary classification output. The sigmoid function was used as the activation function to perform binary classification. We experimented with filter sizes of one, three, and five, and when performing binary classification on a 768-channel input, the parameter count for the 1D convolutional layer was highly efficient: 1,538 parameters with filter size of one and 4,610 with filter sizes of three, and 7,682 with filter sizes of five.

3. Experiment

3.1. Data Description

We collected data from patients who underwent VFSS at Seoul National University Bundang Hospital and from healthy volunteers between October 2021 and February 2023. In the initial cohort, 285 subjects were recruited: 212 normal (88 males and 124 females) and 73 dysphagia-aspirated (54 males and 19 females). The normal and aspiration groups were classified based on PAS scores (normal: PAS 1, aspiration: PAS 5-7) by two clinicians. To enhance data quality, subjects under 40 years of age (78 normal, 1 aspirated) and recordings with poor audio quality (6 normal, 2 aspirated) were excluded, resulting in a final analysis of 198 subjects: 128 normal (41 males, 87 females) and 70 aspirated (52 males, 18 females). This study was approved by the Institutional Review Board of Seoul National University Bundang Hospital (IRB No. B-2109-707-303) and registered with ClinicalTrials.gov (ID. NCT05149976).

To collect pre- and post-swallowing voice data, subjects undergoing the VFSS consumed a standardized bolus (water, semi-blended diet, small fluid, fluid thickening with level 3, liquid food, and Yoplait yogurt) under the researcher’s guidance. The subject phonated [a] for approximately 5 s before and after swallowing, which was recorded at the upper sleeve position using a Sony ICD-TX 660 Recorder. Healthy subjects who did not undergo VFSS followed the same protocol for water and

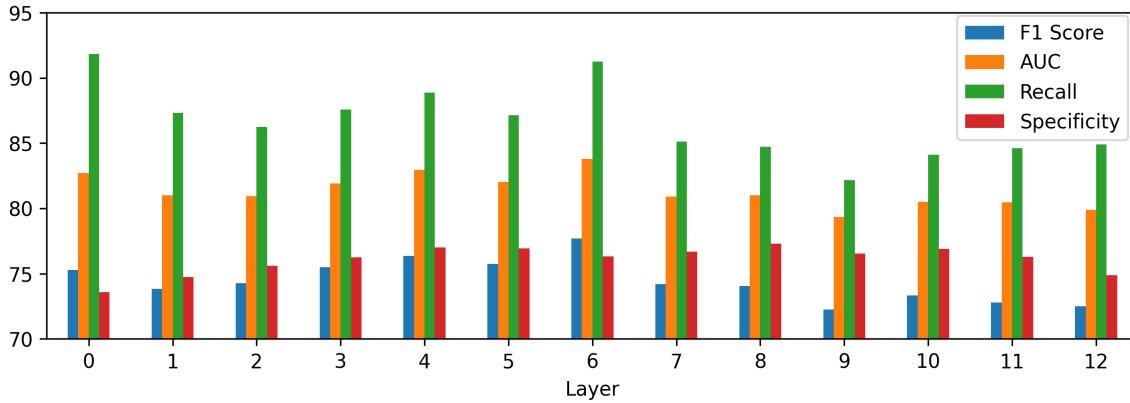


Figure 2: Performance comparison across HuBERT layers. Across all layers, there is a tendency for high recall and low specificity, with this trend being more pronounced in the earlier layers. The 6th layer exhibited the highest F1 score and AUC performance.

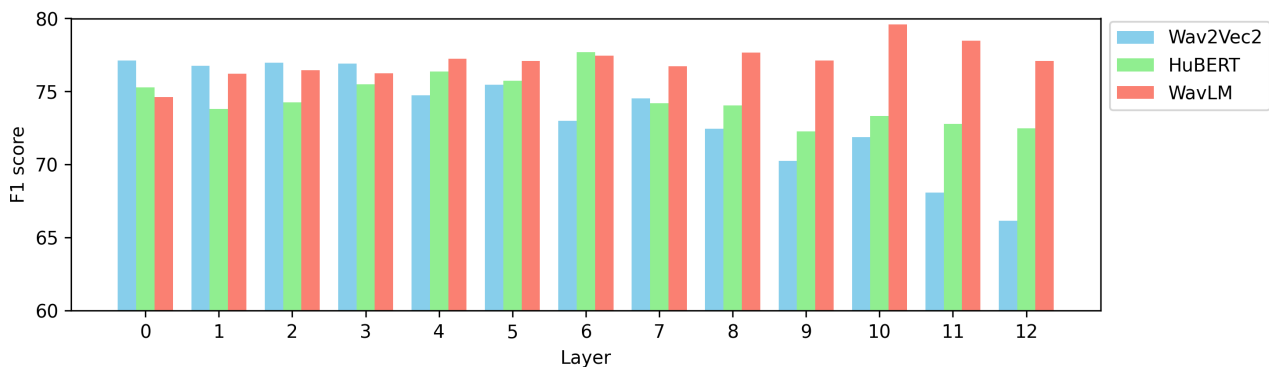


Figure 3: F1 score comparison across layers of SSL models. Each SSL model exhibits a different performance pattern across layers.

were supervised by a clinical nutritionist and an occupational therapist. Owing to accessibility limitations, the subjects used their smartphones for recording. All recordings were truncated to 2 s to standardize the lengths, reduce noise, and maximize usable samples. This process facilitated additional noise reduction from both ends of the recordings and between vocalizations. To address the variability in audio formats, all recording files were standardized to the mp3 format at 64 kbps, considering the potential deployment of the model on mobile devices and the CPU specifications of medical devices. The Sony recorder stereo data were split into two mono files, whereas the smartphone recordings remained in mono format. The resulting data comprised 766 pre-swallowing (male 235, female 531) and 793 post-swallowing (male 240, female 553) samples for the normal group, and 362 pre-swallowing (male 266, female 96) and 578 post-swallowing (male 444, female 134) samples for the aspiration group. To ensure model efficiency, the processed audio data were stored in a hierarchical data format 5 (HDF5). Each HDF5 file contained pre- and post-swallowing mp3 data, normal/aspiration classification labels, and de-identified subject numbers. Pre- and post-swallowing data from the same individual and time points were paired, resulting in 5,333 normal (male 1,556, female 3,777) and 2,968 aspiration (male 2,200, female 768) paired samples.

Model validation employed 10-fold cross-validation to address the limited dataset size. To mitigate individual-specific biases, all data from a single subject were allocated in the same fold. Finally, owing to the significant disparity in the number of samples between the normal and aspiration groups for fe-

male data, random oversampling was performed on the aspiration group to match the number of normal samples in each training dataset-fold during the creation of the HDF5 file.

3.2. Experimental Setup

First, we evaluated the performance of the features extracted from each layer of the SSL model to identify the layer that yielded the best results. For the mel spectrogram, we compared the performance of MobileNet-V3, as in previous studies, with that of a simple predictor. These results were then compared with the results from SSL models to provide a comprehensive evaluation. Subsequently, using the feature that achieved the highest performance, we varied the filter size of the predictor to analyze the effect of model complexity on performance. Finally, we assessed the contributions of the pre- and post-swallowing voice features to the overall model performance.

To evaluate the performance of the model for dysphagia diagnosis, we employed the area under the curve (AUC), F1 Score, recall, and specificity. The threshold for the metrics was adjusted in 0.1 increments, and the value that yielded the highest F1 score was selected. [23] The AUC was employed to assess the overall performance of the model across various classification thresholds, providing a comprehensive measure of the ability of the model to distinguish between classes. The F1 Score, which is the harmonic mean of precision and recall, was used to balance the trade-off between false positives and false negatives, offering insight into the model's effectiveness in both identifying positive cases and avoiding false alarms.

Table 1: Comparison with Baseline

Model	Feature	F1 Score \uparrow	AUC \uparrow	Recall \uparrow	Specificity \uparrow
MobileNet-V3	mel Spec	73.63	80.07	72.76	87.37
1D Conv	mel Spec	69.76	77.04	91.28	62.79
1D Conv	W2V2	77.13	83.76	93.22	74.31
1D Conv	HuBERT	77.68	83.79	91.26	76.31
1D Conv	WavLM	79.60	85.58	91.63	79.54

The model was trained with batch size 256, up to 150 epochs, and learning rate $3e-5$ using the Adam optimizer. Performance was measured as the average over 10 folds. Training was conducted on an NVIDIA RTX A6000 GPU.

3.3. Results

Among the three SSL models, HuBERT exhibited the most pronounced variation in performance across layers, as shown in Figure 2. Recall and specificity did not necessarily follow a strict trade-off relationship, leading to variations in F1 score and AUC. This indicates that the critical information for dysphagia diagnosis differs depending on the layer. Generally, lower layers in SSL models are known to capture rich acoustic information, while deeper layers progressively encode more speaker and linguistic information. [33] Consequently, the dysphagia predictor likely relied on acoustic information when using lower-layer features and leveraged more abstract, high-level representations when using upper-layer features. In the case of HuBERT, the 0th layer exhibited high recall but low specificity. However, as the layers deepened, specificity improved, with the 6th layer achieving the highest F1 score.

Figure 3 compares the F1 scores of each SSL model across different layers. At lower layers, where the predictor primarily utilizes acoustic information for dysphagia detection, performance differences among the three SSL models are minimal. However, as deeper layers are used, the performance gap widens, with WavLM demonstrating a stronger ability to capture high-level information beneficial for dysphagia diagnosis. Notably, Wav2Vec 2.0, the oldest model among the three, shows a decline in performance in deeper layers, whereas WavLM, the most recent model, improves as the layers deepen. This suggests that as SSL models advance, they learn a broader range of high-level features in their upper layers, including those relevant to dysphagia diagnosis.

For each SSL feature, the layer with the highest F1 score was selected: Layer 0 for Wav2Vec 2.0, layer 6 for HuBERT, and layer 10 for WavLM, as listed in Table 1. When using mel spectrogram features, employing a 1d convolutional layer model increases recall but significantly decreases specificity, resulting in a lower F1 score and AUC. Conversely, when switching to SSL features with a single 1D convolutional layer model, the recall performance is maintained, whereas the specificity improves. This suggests that SSL features capture patterns that mel spectrograms cannot, leading to a more precise differentiation between the normal and aspiration groups. Among the SSL features, recall levels were similar, with the most recent model, WavLM, achieving the highest F1 score, AUC, and specificity.

For the experiments on filter size and input data type, the 10th layer of WavLM, which showed the highest performance, was selected. Interestingly, increasing the filter size, thereby increasing the model complexity, had a negative impact on overall performance (Table 2). This suggests that it is more effective for a filter to map each time frame and then simply average along the time axis, rather than attempting to capture the relationships between time frames. This outcome could be partly attributed to the relatively small amount of available data, which makes it

Table 2: Comparison across Convolution Filter Size

Filter	F1 Score \uparrow	AUC \uparrow	Recall \uparrow	Specificity \uparrow
1	79.60	85.58	91.63	79.54
3	78.82	84.68	90.24	79.11
5	78.05	84.16	88.98	79.34

Table 3: Comparison across Input Type

Input	F1 Score \uparrow	AUC \uparrow	Recall \uparrow	Specificity \uparrow
Both	79.60	85.58	91.63	79.54
Post only	73.17	79.90	84.08	75.71
Pre only	70.12	77.83	77.61	77.88

difficult to train larger models effectively. Moreover, the data primarily consisted of single phoneme, which provide limited information from the time axis.

As expected, the overall performance varied by input type, with the highest accuracy observed when using both pre- and post-swallowing voices, followed by post-swallowing only, and then pre-swallowing only (Table 3). However, because prediction accuracy remained surprisingly high even when using only the pre-swallowing or post-swallowing voice, it may be practical to avoid measuring the voice both before and after meals. This finding suggests that individuals can obtain a rough estimate of their condition by measuring their voice just once, even outside a clinical setting, such as at home.

4. Limitation

This study validated the effectiveness of SSL models for voice-based dysphagia diagnosis but has some limitations. Due to privacy regulations, the dataset cannot be publicly shared, limiting data availability for future research. Additionally, the dysphagia predictor used had a simple architecture, and further exploration of model complexity is needed. Despite these limitations, this study demonstrates the potential of SSL-based speech representations for dysphagia diagnosis, providing a foundation for future advancements.

5. Conclusion

We proposed a dysphagia diagnosis model using self-supervised speech representation learning models, specifically Wav2Vec 2.0, HuBERT, and WavLM. When using SSL features, the model achieved a higher performance with a smaller model size compared to the MobileNet-V3 models that use mel spectrograms. Although there was a slight decrease in specificity, recall, which is critical for disease diagnosis, showed a significant improvement, leading to an overall enhancement in model performance. Among the SSL features, 10th layer of the WavLM demonstrated the highest performance. However, increasing the filter size of the convolutional layers did not improve the performance. Additionally, while using both pre- and post-swallowing speech is beneficial, we experimentally confirmed that using either alone yields satisfactory performance, with post-swallowing speech proving to be more crucial for an accurate diagnosis. In the future, we aim to extend the problem to a multiclass classification framework to allow for more detailed categorization based on the symptoms of dysphagia.

6. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C1007780, 30%), and the Institute of Infor-

mation & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2022-II220641, 50%], [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), 5%], and [No. RS-2021-II212068, Artificial Intelligence Innovation Hub, 5%] and the SNUBH Research Fund. (No. 18-2022-0006, 10%)

7. References

- [1] K. Matsuo and J. B. Palmer, "Anatomy and physiology of feeding and swallowing: normal and abnormal," *Physical Medicine and Rehabilitation Clinics of North America*, vol. 19, no. 4, pp. 691–707, 2008.
- [2] A. Sasegbon and S. Hamdy, "The anatomy and physiology of normal and abnormal swallowing in oropharyngeal dysphagia," *Neurogastroenterology & Motility*, vol. 29, no. 11, p. e13100, 2017.
- [3] S. R. Achem and K. R. DeVault, "Dysphagia in aging," *Journal of clinical gastroenterology*, vol. 39, no. 5, pp. 357–371, 2005.
- [4] C. D. Lind, "Dysphagia: evaluation and treatment," *Gastroenterology Clinics*, vol. 32, no. 2, pp. 553–575, 2003.
- [5] A. Ahmed and B. Stacey, "Dysphagia: Aspects of assessment and management for the acute physician," *Acute Med*, vol. 7, no. 3, pp. 107–12, 2008.
- [6] L. Rofes, V. Arreola, J. Almirall, M. Cabré, L. Campins, P. García-Peris, R. Speyer, and P. Clavé, "Diagnosis and management of oropharyngeal dysphagia and its nutritional and respiratory complications in the elderly," *Gastroenterology research and practice*, vol. 2011, no. 1, p. 818979, 2011.
- [7] M. M. B. Costa, "Videofluoroscopy: the gold standard exam for studying swallowing and its dysfunction," pp. 327–328, 2010.
- [8] L. East, K. Nettles, A. Vansant, and S. K. Daniels, "Evaluation of oropharyngeal dysphagia with the videofluoroscopic swallowing study," *Journal of Radiology Nursing*, vol. 33, no. 1, pp. 9–13, 2014.
- [9] K. Helliwell, V. Hughes, C. Bennion, and A. Manning-Stanley, "The use of videofluoroscopy (vfs) and fiberoptic endoscopic evaluation of swallowing (fees) in the investigation of oropharyngeal dysphagia in stroke patients: a narrative review," *Radiography*, vol. 29, no. 2, pp. 284–290, 2023.
- [10] J. C. Borders and D. Brates, "Use of the penetration-aspiration scale in dysphagia research: a systematic review," *Dysphagia*, vol. 35, pp. 583–597, 2020.
- [11] J. C. Rosenbek, J. A. Robbins, E. B. Roecker, J. L. Coyle, and J. L. Wood, "A penetration-aspiration scale," *Dysphagia*, vol. 11, pp. 93–98, 1996.
- [12] A. Wieseke, D. Bantz, L. Siktborg, and N. Dillard, "Assessment and early diagnosis of dysphagia," *Geriatric Nursing*, vol. 29, no. 6, pp. 376–383, 2008.
- [13] C. E. Artiles, J. Regan, and C. Donnellan, "Dysphagia screening in residential care settings: A scoping review," *International Journal of Nursing Studies*, vol. 114, p. 103813, 2021.
- [14] D. Jayatilake, T. Ueno, Y. Teramoto, K. Nakai, K. Hidaka, S. Ayuzawa, K. Eguchi, A. Matsumura, and K. Suzuki, "Smartphone-based real-time assessment of swallowing ability from the swallowing sound," *IEEE journal of translational engineering in health and medicine*, vol. 3, pp. 1–10, 2015.
- [15] J. S. Ryu, S. R. Park, and K. H. Choi, "Prediction of laryngeal aspiration using voice analysis," *American journal of physical medicine & rehabilitation*, vol. 83, no. 10, pp. 753–757, 2004.
- [16] K. W. Dos Santos, E. da Cunha Rodrigues, R. S. Rech, E. M. da Ros Wendland, M. Neves, F. N. Hugo, and J. B. Hilgert, "Using voice change as an indicator of dysphagia: a systematic review," *Dysphagia*, vol. 37, no. 4, pp. 736–748, 2022.
- [17] Y. A. Kang, J. Kim, S. J. Jee, C. W. Jo, and B. S. Koo, "Detection of voice changes due to aspiration via acoustic voice analysis," *Auris Nasus Larynx*, vol. 45, no. 4, pp. 801–806, 2018.
- [18] H.-Y. Park, D. Park, H. S. Kang, H. Kim, S. Lee, and S. Im, "Post-stroke respiratory complications using machine learning with voice features from mobile devices," *Scientific Reports*, vol. 12, no. 1, p. 16682, 2022.
- [19] S. Roldan-Vasco, A. Orozco-Duque, J. C. Suarez-Escudero, and J. R. Orozco-Arroyave, "Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106248, 2021.
- [20] K. López-de Ipiña, P. Calvo, M. Faundez-Zanuy, P. Clave, W. Nascimento, U. Martinez de Lizarduy, D. Alvarez, V. Arreola, O. Ortega, J. Mekyska *et al.*, "Automatic voice analysis for dysphagia detection," *Speech, Language and Hearing*, vol. 21, no. 2, pp. 86–89, 2018.
- [21] H. Kim, H.-Y. Park, D. Park, S. Im, and S. Lee, "Non-invasive way to diagnose dysphagia by training deep learning model with voice spectrograms," *Biomedical Signal Processing and Control*, vol. 86, p. 105259, 2023.
- [22] J.-M. Kim, M.-S. Kim, S.-Y. Choi, and J. S. Ryu, "Prediction of dysphagia aspiration through machine learning-based analysis of patients' postprandial voices," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, no. 1, p. 43, 2024.
- [23] J.-M. Kim, M.-S. Kim, S.-Y. Choi, K. Lee, and J. S. Ryu, "A deep learning approach to dysphagia-aspiration detecting algorithm through pre-and post-swallowing voice changes," *Frontiers in Bioengineering and Biotechnology*, vol. 12, p. 1433087, 2024.
- [24] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.
- [25] H. Song, S. Chen, Z. Chen, Y. Wu, T. Yoshioka, M. Tang, J. W. Shin, and S. Liu, "Exploring wavlm on speech enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 451–457.
- [26] J. Yuan, Y. Shi, L. Li, D. Wang, and A. Hamdulla, "Few-shot keyword spotting from mixed speech," *arXiv preprint arXiv:2407.06078*, 2024.
- [27] K. E. Hajal, A. Kulkarni, E. Hermann, and M. M. Doss, "Ssl-tts: Leveraging self-supervised embeddings and knn retrieval for zero-shot multi-speaker tts," *arXiv preprint arXiv:2408.10771*, 2024.
- [28] 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.
- [29] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, 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.
- [30] 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.
- [31] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
- [32] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [33] 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.