Dysarthric Speech Recognition Fusing Large Pre-trained Model Extracted Acoustic Features with Articulatory Data

Dysarthric Speech Recognition Fusing Large Pre-trained Model Extracted Acoustic Features with Articulatory Data

by

Xinrui Xu

to obtain the degree of Master of Science at the Delft University of Technology,
supervised by Dr. O. Scharenborg and Dr. Z. Yue, to be defended publicly in March 2025.

Thesis committee

Dr. O. Scharenborg	Technische Universiteit Delft
Dr. Z. Yue	Technische Universiteit Delft
Dr. J. Sun	Technische Universiteit Delft

Contents

1

Introduction

1.1. Motivation

Dysarthria is a speech disorder due to neural damage of the motor component of the motor–speech system. In most cases, dysarthria is caused by brain damage [39], which can result from congenital conditions like cerebral palsy or muscular dystrophy [29]. It can also be triggered by certain acquired factors such as stroke, brain injury, or Parkinson's disease [30]. Individuals with dysarthria often struggle with producing spoken sounds, including difficulties with articulation, slurred speech, or irregularities in the clarity, pitch, and speed of their speech [29].

The advancement of deep learning technologies has significantly improved automatic speech recognition (ASR) systems. However, most existing ASR systems are trained on typical speech datasets [13]. But for people with dysarthria who have difficulty communicating with others due to slow, unclear, or fluctuating speech speed, the performance of existing ASR systems is poorer than ideal [33]. The poor performance of ASR in recognizing dysarthric speech can severely impact daily communication, making it difficult for people with dysarthria to interact effectively [1].

The speech characteristics of speakers with dysarthria differ significantly from typical speech, including fluency, and articulatory accuracy which pose additional challenges for recognition by ASR systems. To further optimize the ASR system so that it can better recognize people with dysarthria, it's necessary to propose a more adapted ASR system that can recognize dysarthric speech.

The primary challenge lies in the variability of dysarthric speech among individuals. This variability makes it more difficult to learn the model accurately, as the speech patterns vary a lot in severity and type [49]. To address this challenge, researchers have proposed several strategies. One approach is data enhancement, such as generating training data for dysarthric speech from typical speech by using adversarial training [23] or using temporal and speed modifications of typical speech [42]. Another state-of-art approach is using multiple modalities, which combines data from other modalities like facial images or articulator movements, with speech data to improve the performance of the ASR system [53]. This method can complement the acoustic information and improve ASR performance. In the case of dysarthria, changes in muscle activity of articulators result in unique speech patterns that are critical for the ASR system to recognize [24]. Articulatory features can help to recognize these differences and provide information about the physical process of speech production [41], which can be combined with acoustic features to provide richer and more multidimensional data for speech recognition systems, thus improving the performance of the ASR [55]. Some previous works [55, 19] have proposed to combine acoustic features with articulatory features. These works have shown that automatic dysarthric speech recognition(ADSR) systems can better recognize dysarthric speech by training with articulatory features. This thesis will focus on exploring the impact of combining articulatory information with acoustic features to improve dysarthric speech recognition.

Additionally, with the development of large pre-trained acoustic models such as Whisper [34], WavLM [6], and Hubert [20], some researchers have proposed using features extracted by these models for various tasks such as speech classification or typical speech recognition. However, to my best knowledge, few works have explored the combination of features extracted from large pre-trained models with articulatory features. Instead, most previous studies rely on traditional acoustic features like mel filter bank (FBank) [9]. To bridge this gap, this work will combine features extracted from large pre-trained models with articulatory features.

1.2. RESEARCH QUESTIONS

This thesis will approach the dysarthric speech recognition problem by exploring the integration of articulatory features with features extracted from large pre-trained models such as Whisper and WavLM. The aim is to enhance the accuracy of dysarthric speech recognition by leveraging the complementary strengths of both articulatory and acoustic features. Various fusion methods will be investigated to determine how best to combine these feature types to improve ASR performance. The main research question will thus be:

- **RQ1**: How effective are articulatory features in enhancing dysarthric speech recognition when combined with features extracted from large pre-trained models?
- **RQ2**: How does the effectiveness of combining articulatory features with acoustic features vary across different severity levels of dysarthria?
- **RQ3**: What fusion methods can better utilize the feature information from both acoustic and articulatory features?

1.3. OUTLINE

This thesis is organized into several chapters. Chapter 1 introduces the research, including the motivation and research questions. Chapter 2 reviews related work, while Chapter 3 details the methodology. In Chapter 4, I will describe the experiment setups. Chapter 5 discusses the results of different models. Finally, Chapter 6 concludes with key findings and future research directions.

2

Literature Review

This chapter explores related research on incorporating articulatory features in speech recognition, utilizing features extracted from large pre-trained models, and implementing multimodal fusion techniques.

2.1. RELATED WORK

2.1.1. ARTICULATORY FEATURES IN ASR MODELS

Articulatory features, derived from the physical movements of speech articulators, can provide complementary information to acoustic features in automatic speech recognition (ASR) tasks, as demonstrated in several previous studies. Early research in ASR explored articulatory features and highlighted their potential to enhance recognition performance. For instance, Wrench and Richmond [47] proposed a dynamic Bayesian network (DBN) model that integrated articulatory information with acoustic signals (MFCCs), resulting in a reduction in word error rate (WER) by approximately 10%. Later, Markov et al. [28] introduced a hybrid model combining HMMs and Bayesian Networks (BN), which integrated both articulatory and acoustic (MFCCs) features by modeling the probabilistic dependencies between them. In addition to the position of the articulators, Markov's model also incorporated velocity and acceleration data of articulators.

While early research successfully proved the effectiveness of combining articulatory features with acoustic data, these approaches often relied on traditional models like HMMs and DBNs, which have limitations in capturing complex dependencies and nonlinear relationships between features. With the development of deep learning, advanced neural network architectures, such as RNN [46], Transformer [43], and Conformer [14] have a better performance in the field of ASR. Recent studies have leveraged these architectures to more effectively integrate articulatory and acoustic features. Since it is difficult to collect articulatory data, many studies have explored acoustic-to-articulatory mapping (AAM) to estimate articulatory data from acoustic features. For example, Leonardo Badino et al. [3] used AAM to transform acoustic features into articulatory features and leverage these features to enhance the performance of hybrid DNN-HMM models. Mitra et al. [32] managed to use deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) to map speech data to the corresponding articulatory space and jointly learn acoustic and articulatory spaces to improve the performance of ASR.

2.1.2. ARTICULATORY FEATURES IN ADSR MODELS

In addition to their application in ASR for typical speech, many works have explored the use of articulatory features in dysarthric speech recognition. Emre Yılmaz et al. [52] employed vocal tract constriction variables (TVs) as articulatory features, which describe the degree of contraction of the vocal tract and the location of the contraction. This study concatenated TVs with Fbank features as input to their acoustic model. The fusion of articulatory features led to a lower WER on the CHASING01 and COPAS test datasets. Zhengjun Yue et al. [55] explored multi-modal speech recognition for dysarthric speakers by fusing acoustic and articulatory features. It used a multi-stream architecture with CNN, RNN, and fully connected layers, allowing for processing each feature type separately before combining them. This work also explored three multimodal feature fusion methods, categorized by the layer at which acoustic and articulatory features(MFCCs) are integrated. More recently, Hsieh et al. [19] proposed a method that used curriculum learning to improve ADSR performance, where the model first learns easier (closer to typical speech) samples, then progressively adapts to more challenging (severely dysarthric) speech. This study integrated 80-dimensional Fbank features with articulatory features, along with speaker identity and speech intelligibility embeddings, as input to the model. This combination resulted in a lower WER, demonstrating the effectiveness of articulatory features for ADSR.

2.1.3. Large Pre-trained Model Extracted Features for Dysarthria Speech Recognition

In addition to studies incorporating articulatory features, recent research has also explored the use of large pre-trained models for pathological speech recognition. These large pre-trained acoustic models have shown powerful capabilities in capturing details from speech signals, even in atypical speech. Yaroslav Getman et al. [11]proposed a method using features extracted by Wav2vec2 [4] to enhance ASR performance for children with speech sound disorders (SSD). Their findings showed that features extracted by Wav2vec2 achieved a lower WER compared to traditional acoustic feature MFCCs. Another more recent work [37]employed a pre-trained Whisper [34] model to extract audio features and fed these features into acoustic models, such as LSTM [17], Bi-LSTM [40], and Bi-GRU [7] for speech recognition. Experimental results show that Whisper features can maintain high recognition accuracy in noisy environments, outperforming other commonly used large pre-trained models including Hubert [20] and Wav2vec2. Shujie Hu et al. [22] explored the combination of self-supervised learning (SSL) models (e.g., Wav2vec2.0, HuBERT , etc.) with the TDNN and Conformer ASR systems to improve the recognition performance of aphasic and elderly speech. Experimental results on datasets such as UASpeech and TORGO showed that the method can reduce the word error rate (WER) and the character error rate (CER), and achieve an improvement in the challenging dysarthric speech data

2.1.4. Multimodal Fusion in ASR Tasks

Besides the concatenation method discussed in the previous sections, which has been utilized in many works [52, 19, 55], other fusion methods are employed to integrate acoustic features with different modalities. Pingchuan Ma et al. [**<empty citation>**] proposed to fuse audio and visual features by first extracting them separately using ResNet-18-based front-ends (3D CNN for visual and 1D CNN for audio). Then the extracted embeddings are concatenated and fused via an MLP module, projecting them into a shared latent space. Gaopeng Xu et al. [50] proposed a cross-attention mechanism for audio-visual fusion. In this study, acoustic and visual features are first independently encoded before audio-visual fusion. Then, cross-attention is applied, where audio features serve as the query and visual features (e.g., lip movements) act as key/value, allowing the model to focus on the most relevant visual information for speech recognition. [15] also employed a cross-attention-based multi-modal fusion method to combine visual and acoustic features, which can capture contextual relationships between them. Compared with the previous method which only employs acoustic features as a query, in this method, each modality alternates as the query, while the other serves as the key and value in this approach. Inspired by this, a similar cross-attention-based method is adopted in this study to fuse articulatory and acoustic features.

In conclusion, previous research on dysarthric speech recognition has explored the effectiveness of using articulatory features and large pre-trained models separately. While articulatory features improve recognition by adding speech production information, large pre-trained models capture complex acoustic patterns. However, the combination of these approaches remains under-explored. This study addresses this gap by integrating articulatory and large pre-trained model extracted features to enhance ASR performance for dysarthric speech.

3

METHODOLOGY

This chapter briefly overviews several mainstream large pre-trained models and explains the process of extracting acoustic features, including Fbank features and features extracted by large pre-trained models. It also describes the articulatory data from the Torgo dataset. Additionally, this chapter discusses multimodal fusion strategies, such as concatenation and cross-attention-based methods. Furthermore, this chapter introduces sequence-to-sequence ASR models and explains the Conformer encoder and Transformer decoder architecture. Additionally, it includes a description of t-SNE analysis.

3.1. Acoustic Features

In ASR systems, the initial step is feature extraction, where the aim is to capture the essential components of the audio signal. These features are used to characterize the acoustic structure of speech by analyzing both the frequency and time domain properties of the speech signal.

3.1.1. Fbank

The Fbank feature is one of the traditional acoustic features that play an important role in the field of speech recognition due to its direct representation of the power spectrum [48]. The Fbank generates feature vectors representing frequency distributions by applying triangular filters on the Mel frequency scale. FBank features have been commonly used as baseline representations in some studies related to dysarthric speech recognition tasks, such as [16, 54, 56]. Therefore, in this study, Fbank was also used as a baseline acoustic feature to compare it with features extracted from large pre-trained models.

3.1.2. Large Pre-trained Models Extracted Features

In this research, the following large pre-trained models were employed: HuBERT [20], Wav2vec [4], Whisper [34], and WavLM [6]. These models are all based on Transformer architectures which are highly effective at capturing both spectral and temporal dependencies in the audio signal and employ self-supervised learning to extract high-level speech features from unlabeled audio data that are suitable for speech recognition (except Whisper, which is trained on large-scale labeled data). These large pre-trained models have shown strong performance on dysarthric speech and are widely adopted in ADSR studies [21, 11, 36]. Therefore, this research also employs them to compare with Fbank features. Below is a brief introduction to these large pre-trained models.

- **Wav2Vec 2.0** [4]: Wav2Vec 2.0 is a self-supervised speech representation model developed by Meta. It learns speech features directly from raw audio waveforms without relying on extensive labeled datasets. The model is trained by masking parts of the audio input and learning to predict the missing segments, enabling it to capture rich acoustic features. By learning to predict masked audio segments, Wav2Vec 2.0 can capture contextual information, leading to richer, more robust speech representations.

- **HuBERT** [20]: Hubert is also a self-supervised speech representation model developed by Meta. It builds on the strengths of previous models like Wav2Vec 2.0 but introduces a novel approach to learning speech representations. HuBERT uses a method that clusters audio into discrete hidden units and then predicts these units. This approach helps the model learn more structured and robust representations of speech. Also by predicting masked parts of speech based on surrounding audio, HuBERT can capture longer-range dependencies more effectively.

- **Whisper** [34]: Whisper is a large pre-trained acoustic model released by OpenAI in 2022. OpenAI trained Whisper with 680,000 hours of supervised multilingual (98 languages) and multitask data collected from the web. Unlike models like wav2vec, whisper utilizes weakly supervised training, which enables it to perform direct multitask learning without requiring task-specific fine-tuning. This training approach also allows Whisper to leverage a smaller amount of labeled data while learning richer and more generalized feature representations from the data.

- **WavLM** [6]: WavLM is a self-supervised pre-trained model for speech processing developed by Microsoft. One of the key innovations in WavLM is the use of masked speech prediction and denoising during pre-training. This approach trains the model to predict masked segments of speech while also handling various noise conditions, which helps the model learn more robust and generalized speech representations.

In this work, I leverage Hugging Face's interfaces to extract acoustic features using the mentioned large pre-trained models. For each large pre-trained model, Hugging Face offers multiple versions based on model size. To capture richer information from the speech signal, this study chose the large versions, specifically HuBERT-large, WavLM-large, Whisper-large-v3, and Wav2vec2-xlsr-53. These large pre-trained models are all based on Transformer architectures and composed of both an encoder and a decoder. The features are all extracted from the final layer of each model's encoder, as this layer can capture high-level representations [8]. The extracted features vary in dimensionality depending on the model. Specifically, as shown in the table below, HuBERT-large, Wav2vec2-xlsr-53, and WavLM-large generate features with 1024 dimensions, whereas Whisper-large-v3 produces features with 1280 dimensions. Here, Nframe denotes the number of frames in a speech signal. Due to the downsampling mechanism in these large pre-trained models, the number of extracted feature frames is reduced to half of the original audio.

Large Pre-trained Models	**Parameters**	**Extracted Feature Dimension**
Whisper-large-v3	1550 M	$1 \times \text{Nframe}/2 \times 1280$
WavLM-large	317 M	$1 \times \text{Nframe}/2 \times 1024$
HuBERT-large	316 M	$1 \times \text{Nframe}/2 \times 1024$
Wav2vec2-xlsr-53	316 M	$1 \times \text{Nframe}/2 \times 1024$

Table 3.1: Comparison of Large Pre-trained Models, Parameters, and Extracted Feature Dimensions

Take Whisper as an example, the feature extraction pipeline is illustrated in Figure3.1. With all the parameters frozen, the speech signal was fed into Whisper and only the outputs of the last hidden layer of encoders were used as input for our acoustic model.

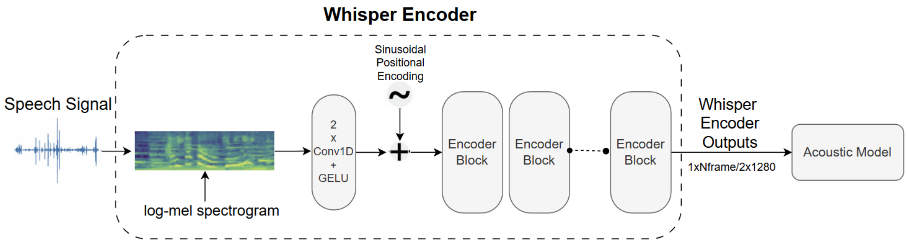


Figure 3.1: Feature Extraction Pipeline of Whisper Encoder

3.2. ARTICULATORY FEATURES

Articulatory Features describe the primary actions of speech organs during speech production. Through articulatory features, a correspondence can be established between the speech signal and key articulatory units [31].

In this study, articulatory features are used to capture the kinetic information of articulators such as the lips and tongue, providing complementary information to acoustic features. The articulatory data used in the experiments is from the TORGO dataset [38], which uses a 3D AG500 electro-magnetic articulograph (EMA) system to record the articulatory movement data simultaneously with the acoustic data. The system is automatically calibrated and can record 3D movements of both internal and external articulators. As shown below, the collection system uses 12 sensor emitters to generate alternating electromagnetic fields and measures the movements of the tongue, chin, and lips. Sensor coils are attached to the tongue tip, middle, and back, as well as the upper and lower lips, lower incisor, and mouth corners to track their movements, as shown in Table 3.2.

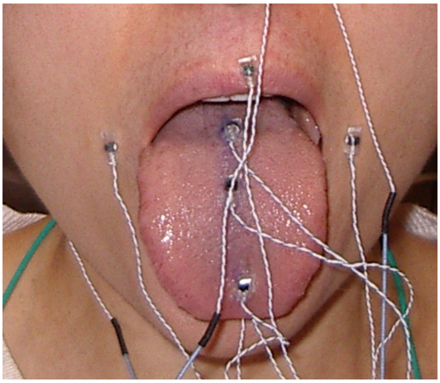


Figure 3.2: Electromagnetic Coil Placement for Articulatory Data Collection from the TORGO Dataset [38]

Index	Position in the Vocal Tract
1	Tongue back (TB)
2	Tongue middle (TM)
3	Tongue tip (TT)
4	Forehead
5	Bridge of the nose (BN)
6	Upper lip (UL)
7	Lower lip (LL)
8	Lower incisor (LI)
9	Left lip
10	Right lip
11	Left ear
12	Right ear

Table 3.2: Sensor coil positions in the vocal tract. [38]

During the collection of EMA data, the subject's head is free to move, and the system connects the sensors via a lightweight cable that does not interfere with the subject's freedom of movement in the EMA device. The articulatory data is stored as a time series, and the positional coordinates of the transducers (typically the X, Y, and Z axes) are recorded at each time point, capturing the three-dimensional trajectory of the motion of these articulatory organs.

As shown in the previous work [55] where lip and tongue EMA data have proven helpful for ADSR, this study also employed EMA data collected from lip and tongue sensors as articulatory features.

3.3. Multimodal Fusion Strategy

Since acoustic features alone may not fully capture the characteristics of dysarthric speech during training, I explored several multimodal fusion strategies to incorporate articulatory features with acoustic features. In this work, both concatenation and cross-attention-based methods were employed. Since the EMA data (200 Hz) and acoustic data (100 Hz) have different frame rates, the EMA data was downsampled for better alignment before fusion.

Concatenation is one of the early fusion strategies, where multiple features are combined or directly concatenated at the input stage of the model, which allows the model to process multimodal information from the very beginning [35]. In this work, early fusion was utilized by directly concatenating acoustic and articulatory features at the input stage. Mathematically, given the acoustic feature $\mathbf{a}_t \in \mathbb{R}^{d_A}$ and the articulatory feature $\mathbf{m}_t \in \mathbb{R}^{d_M}$ at each time step t, the concatenated feature vector is computed as:

$$\mathbf{x}_t = [\mathbf{a}_t; \mathbf{m}_t] \in \mathbb{R}^{d_A + d_M} \tag{3.1}$$

where $\mathbf{a}_t$ and $\mathbf{m}_t$ are the frame-level acoustic and articulatory feature vectors, respectively, and $[\mathbf{a}_t; \mathbf{m}_t]$ denotes concatenation along the feature dimension.

In this research, cross-attention [43] was used as an alternative fusion strategy, trying to further enhance the integration of multimodal information. The cross-attention mechanism consists of a series of attention layers, where each layer takes one modality (e.g., acoustic features) as the query and the other modality (e.g., articulatory features) as the key and value. The attention scores are then calculated by measuring the relevance between the queries and keys, typically using a dot product followed by normalization, such as softmax. These scores are then used to weight the value vectors, allowing the model to focus on the most relevant articulatory information. Finally, the weighted values are combined to form a fused feature that integrates both acoustic and articulatory information. This mechanism enables the model to dynamically weigh and focus on specific parts of the input sequence from each modality.

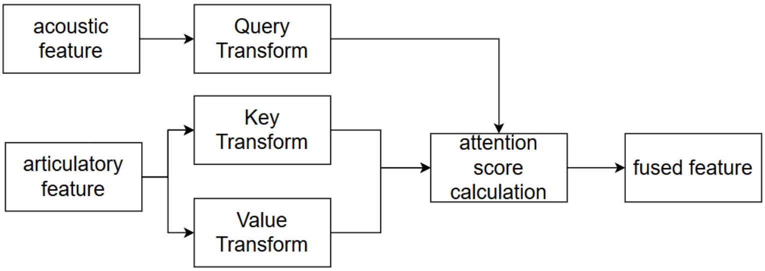


Figure 3.3: Achitecture of Cross Attention Mechanism

Mathematically, given the acoustic features $\mathbf{A} \in \mathbb{R}^{T\times d_A}$ and the articulatory features $\mathbf{M} \in \mathbb{R}^{T\times d_M}$, the cross-attention mechanism computes the fused representation as follows:

$$\mathbf{Q}_A = \mathbf{W}_Q^A \mathbf{A}, \quad \mathbf{K}_M = \mathbf{W}_K^M \mathbf{M}, \quad \mathbf{V}_M = \mathbf{W}_V^M \mathbf{M} \tag{3.2}$$

$$\mathbf{S} = \frac{\mathbf{Q}_A \mathbf{K}_M^\top}{\sqrt{d_k}} \tag{3.3}$$

$$\mathbf{A}_{\text{weights}} = \text{softmax}(\mathbf{S}) \tag{3.4}$$

$$\mathbf{X} = \mathbf{A}_{\text{weights}} \mathbf{V}_M \tag{3.5}$$

where $\mathbf{W}_Q^A \in \mathbb{R}^{d_k\times d_A}$, $\mathbf{W}_K^M, \mathbf{W}_V^M \in \mathbb{R}^{d_k\times d_M}$ are learnable projection matrices, and d_k is the dimension of the key vectors used for scaling.

Additionally, inspired by the work of [15], which employs two cross-attention layers for visual-audio fusion, this research adopted a similar approach to fuse articulatory features. The architecture of the improved cross-attention is shown below:

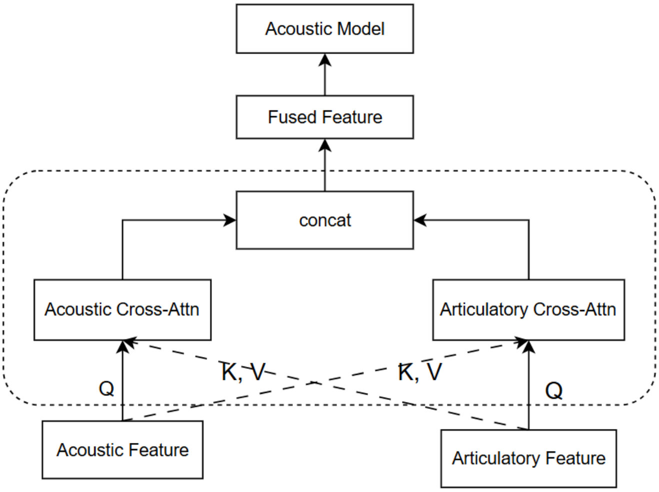


Figure 3.4: Architecture of Improved Cross Attention Mechanism. In this mechanism, each modality attends to the other by using its own features as the query (Q), while obtaining the key (K) and value (V) from the other modality.

In this method, the acoustic features go through an acoustic cross-attention layer, using articulatory features as keys and values. Simultaneously, the articulatory features are processed through an articulatory cross-attention layer, with acoustic features as keys and values. The outputs from both layers are then concatenated to form a fused feature before passing to the acoustic model. This bidirectional cross-attention mechanism facilitates effective information exchange and enhances feature complementarity. Here's the math expression for this improved method, the equation is the same as the previous method when acoustic features serve as query and articulatory features serve as keys and values. Given the acoustic features $\mathbf{A} \in \mathbb{R}^{T \times d_A}$ and the articulatory features $\mathbf{M} \in \mathbb{R}^{T \times d_M}$, the bidirectional cross-attention mechanism is computed as follows:

(1) Acoustic-to-Articulatory Cross-Attention:

$$\mathbf{Q}_A = \mathbf{W}_Q^A \mathbf{A}, \quad \mathbf{K}_M = \mathbf{W}_K^M \mathbf{M}, \quad \mathbf{V}_M = \mathbf{W}_V^M \mathbf{M} \tag{3.6}$$

$$\mathbf{S}_A = \frac{\mathbf{Q}_A \mathbf{K}_M^\top}{\sqrt{d_k}} \tag{3.7}$$

$$\mathbf{A}_{\text{weights}} = \text{softmax}(\mathbf{S}_A) \tag{3.8}$$

$$\mathbf{X}_A = \mathbf{A}_{\text{weights}} \mathbf{V}_M \tag{3.9}$$

(2) Articulatory-to-Acoustic Cross-Attention:

$$\mathbf{Q}_M = \mathbf{W}_Q^M \mathbf{M}, \quad \mathbf{K}_A = \mathbf{W}_K^A \mathbf{A}, \quad \mathbf{V}_A = \mathbf{W}_V^A \mathbf{A} \tag{3.10}$$

$$\mathbf{S}_M = \frac{\mathbf{Q}_M \mathbf{K}_A^\top}{\sqrt{d_k}} \tag{3.11}$$

$$\mathbf{M}_{\text{weights}} = \text{softmax}(\mathbf{S}_M) \tag{3.12}$$

$$\mathbf{X}_M = \mathbf{M}_{\text{weights}}\mathbf{V}_A \tag{3.13}$$

(3) Feature Concatenation:

$$\mathbf{X}_{\text{fused}} = [\mathbf{X}_A; \mathbf{X}_M] \in \mathbb{R}^{T\times(d_A+d_M)} \tag{3.14}$$

where $\mathbf{W}_Q^A, \mathbf{W}_K^A, \mathbf{W}_V^A \in \mathbb{R}^{d_k\times d_A}$ and $\mathbf{W}_Q^M, \mathbf{W}_K^M, \mathbf{W}_V^M \in \mathbb{R}^{d_k\times d_M}$ are learnable projection matrices, and d_k is the dimension of the key vectors used for scaling. This bidirectional cross-attention mechanism facilitates effective information exchange and enhances feature complementarity.

3.4. ASR MODEL

This study employed a Seq2seq acoustic model for speech recognition tasks. One of the major advantages of the Seq2seq model is its ability to directly convert input audio sequences into target text transcriptions. The Seq2Seq model usually consists of an encoder and a decoder. The encoder's role is to transform the input sequence of audio features into high-dimensional hidden representations. Commonly used encoders include Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), as well as Transformer and Conformer models that use self-attention mechanisms.

The role of the decoder is to generate the target output sequence from the high-dimensional representations produced by the encoder. By using attention mechanisms, the decoder can focus on the most relevant parts of the input sequence while producing the output. Then a beam search [26] algorithm is used to improve the quality of the generated text. This algorithm keeps several possible output paths and ranks them based on a scorer, which usually considers the generation probability, language model score, and a length penalty. At last, the path with the highest score is chosen as the final output.

3.5. CONFORMER ENCODER

In this work, the Conformer [14]was employed as the Seq2seq model encoder, to compare performance between acoustic-only and multimodal models. In the field of ADSR, Conformer has been adopted in many works [2, 44]. Additionally, Conformer has shown robustness in noisy speech environments [51]. Therefore, this research also employed the Conformer architecture. The architecture of the Conformer encoder is illustrated below. The left part of the figure represents the preprocessing pipeline for the speech signal, while the right part shows the structure of the Conformer block.

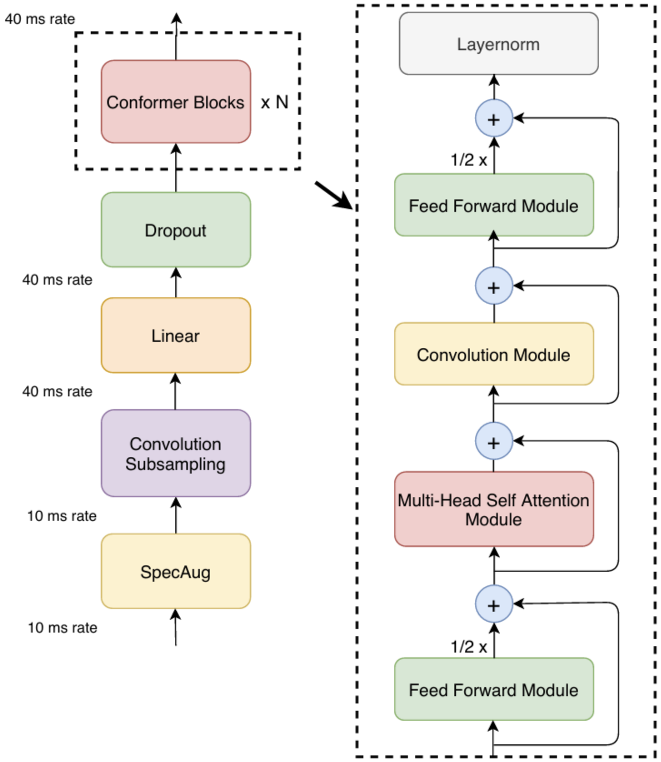


Figure 3.5: Conformer encoder model architecture [14]

As shown in the figure, after data augmentation, a convolutional downsampling module is then applied to reduce the sequence length by downsampling the time dimension. Next, the downsampled features will go through a dropout layer to mitigate the overfitting problem. After the preprocessing steps, the acoustic features will be fed into several conformed blocks(right part). It consists of a feed-forward module that transforms features through a residual connection, followed by a multi-head self-attention module to capture global dependencies, and then another residual connection. Next, a convolution module models local temporal features, again fused with a residual connection. Another feed-forward module applies feature transformations, and at last, the data is fed into a layer normalization layer that balances the scales of different features.

Compared to traditional Transformer [43] architecture which lacks specialized modeling of local temporal features, Conformer introduces a convolution module to solve the problem. The convolution module can capture local temporal dependencies in speech signals, complementing the global modeling capabilities of the self-attention mechanism. This structure effectively captures both local and global information, making it suitable for ASR tasks.

3.6. TRANSFORMER DECODER

This study utilized the Conformer model from the SpeechBrain toolkit, which employs Transformer as its decoder. The Transformer decoder generates tokens in an autoregressive way, meaning that each decoding step depends on the previously generated tokens. The architecture of the Transformer decoder is shown in Figure 3.6. First, the decoder takes the Conformer encoder outputs as input, followed by a layer normalization. Then a masked multi-head attention mechanism is employed to ensure each token can only attend to previous tokens to keep causality. Next, a multi-head mechanism allows the decoder to focus on relevant information from the encoder output, followed by another layer normalization. The transformed representations pass through a feed-forward module and a layer normalization. Finally, the processed output will be fed into a softmax layer, calculating the probability distribution over the target vocabulary.

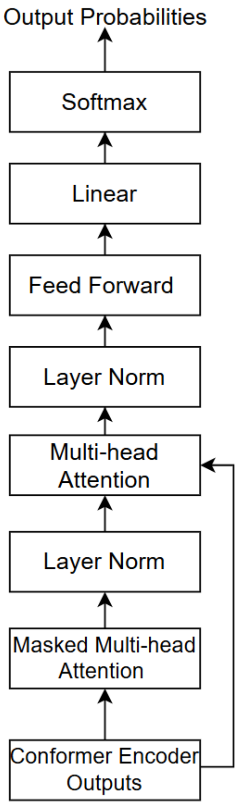


Figure 3.6: Transformer decoder architecture

3.7. T-SNE ANALYSIS

T-SNE [27] (t-distributed Stochastic Neighbor Embedding) is a nonlinear method for dimensionality reduction of high-dimensional data, which is suitable for visualizing high-dimensional data into 2D or 3D space. t-SNE's main advantage is that it preserves the structure of the data points in local neighborhoods well [5], and thus is suitable for demonstrating the distribution of data.

In this study, before t-SNE projection, principal component analysis (PCA) [18] was first employed to reduce the embedding dimension to 50 in order to speed up the computation. Then t-SNE was used to project these embeddings into a lower-dimensional space. By comparing the centroid distances of dysarthric and typical speech embeddings across different severity levels, t-SNE helps analyze the local structure of the data. It reveals variations in speech-embedded features under different feature combinations (e.g., acoustic only vs. acoustic combined with articulatory features) and provides evidence for understanding the impact of articulatory features on the ADSR task.

4

EXPERIMENTS

This chapter presents the experimental design and implementation, including the dataset used, data preprocessing methods, feature extraction processes, and experimental setup. Furthermore, it outlines the evaluation metrics and introduces the p-value as an analysis metric to validate the effectiveness of the proposed approach.

4.1. DATASET

To address the research questions, the dataset must include both dysarthric speech data and synchronized articulatory data, which the TORGO dataset provides. The TORGO dataset contains 23 hours of English speech data and transcripts from 8 speakers with dysarthria (5 males and 3 females) due to cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), as well as 7 control speakers (4 males and 3 females) without speech impairments. The dataset includes both typical and dysarthric speech samples, covering a wide range of dysarthria severity levels (severe, m/s, moderate, and mild), where m/s means the transitional level between moderate impairments and severe limitations in speech abilities. All the audios are recorded by two types of microphones, one is the Acoustic Magic Voice Tracker array microphone, and the other is a head-mounted electret microphone.

However, not all utterances in TORGO have a corresponding .pos file that records the temporal positions of the sensors. To address the RQs, it's necessary to filter the dataset to retain only utterances containing both speech signal and articulatory data. After filtering, a total of 12,125 utterances that meet the criteria are retained. Notably, there are 6,416 .pos files, because each .pos file corresponds to one or two audio files recorded by different microphones.

The filtered subset comprises utterances from both dysarthric and normal speakers. Specifically, there are 3,487 utterances from dysarthric speakers and 8,638 from normal speakers. Detailed information including the number of utterances for each dysarthric and normal speaker is provided below.

Speaker	Severity Level	Gender	Number
M01	Severe	Male	182
M02	Severe	Male	318
M04	Severe	Male	540
M05	M/S	Male	461
F03	Moderate	Female	701
F04	Mild	Female	487
M03	Mild	Male	798

Table 4.1: Dysarthric Speaker information including severity level, gender, and number of utterances

4.2. DATA PREPROCESSING AND FEATURE EXTRACTION

4.2.1. SPEED PERTURBATION

Before the training process, speed perturbation was applied as a data augmentation technique to both the Fbank features and the features extracted from large pre-trained models. Speed perturbation is a widely used technique in the field of speech recognition, enhancing the robustness and generalization ability of the model, and thereby improving the model performance [25]. By modifying the playback speed of the audio (e.g., to 90%, 100%, and 110% of the original speed), this method triples the amount of data available for training.

Speaker	Gender	Number
FC01	Female	294
FC02	Female	1925
FC03	Female	1501
MC01	Male	1438
MC03	Male	1136
MC04	Male	1239
MC02	Male	1105

Table 4.2: Typical speaker information including gender and number of utterances

The expanded dataset helps make the model more robust by making small changes in speech speed, which can improve its ability to handle different speech patterns [10]. After speed perturbation, the playback speed and duration of the audio changes, e.g., when the audio speeds up, the duration shortens, while basic speech characteristics such as pitch and speaker timbre remain unchanged.

4.2.2. Acoustic Feature

In this study, two types of acoustic features were applied: traditional Fbank features and features extracted from large pre-trained models. The details of these features and the extraction processes are described as follows:

- **Fbank:** For the baseline model, 80-dimensional Fbank features were extracted using SpeechBrain's Fbank feature extraction function. The feature extraction process involved setting a window length of 25 ms and a hop length of 10 ms, with a sampling rate of 16 kHz.

- **Large pre-trained model extracted features:** To obtain these features, Hugging Face's last_hidden_state() function was utilized, which extracts the output from the final hidden layer of the encoder in each model. Each feature set is then fed into the ASR model.

4.2.3. Articulatory Feature

For articulatory features, TORGO has a .pos file that records the movements of the articulators. The shape of data in .pos file is Ntime × 7 × 12, where Ntime indicates the number of temporal sample points, 7 represents variables (x, y, z, phi, theta, exit-flag, rms-value), and 12 corresponds to the 12 sensor channels. In this work, only the (x, y, z) coordinates were utilized.

To process the raw EMA data, which includes the (x, y, z) coordinates of tongue and lip sensors, this research followed the methodology and used the code provided by the paper[55]. The processing involved several steps to preprocess articulatory features:

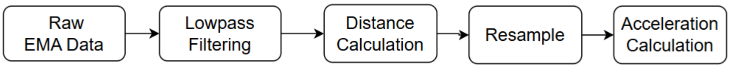


Figure 4.1: Pipeline of Processing Articulatory Features

Lowpass Filtering: First, a lowpass filter is applied to the raw EMA data to remove high-frequency noise, preserving essential movement patterns in the articulators.

Distance Calculation: Then I calculated the Euclidean distances between lip sensors to capture meaningful articulatory features, containing the relative movements and positions of the articulators.

Resample: By using SciPy's signal resample function, the EMA data is resampled to match both the original frame rate and adjusted frame rates for speed perturbation (90% and 110%). This step enables the concatenation and alignment of articulatory and acoustic features.

Acceleration Calculation: At last, the compute_delta() function from SpeechBrain is applied to compute both delta and delta-delta features on the calculated distances. The delta-delta computation captures not only the first-order dynamic changes but also the acceleration of articulatory movements, providing richer temporal information for analysis.

4.3. MODEL ACHITECTURE

For all the experiments, the Conformer model from the SpeechBrain recipe served as the model architecture. The input features are first passed through a normalizer and then through a CNN front-end with 2 convolutional blocks before entering the Conformer encoder. Each convolutional block has a 3x3 kernel size with strides of 2, providing initial feature extraction and downsampling to manage temporal dimensions. The encoder consists of 12 layers, with each layer employing a hidden dimensionality of 144, 4 attention heads, and feed-forward layers with a size of 1024. The decoder is based on a 4-layer Transformer architecture using a CTC/Attention decoding strategy. Additionally, decoding involves using beam search with a beam size of 10 for validation and 66 for testing, relying only on CTC scoring. The model also employs a unigram token vocabulary with 500 output neurons for token prediction.

4.3.1. BASELINE MODEL

In the acoustic-only monomodal model, the ASR system focuses only on acoustic features. These features are derived either from traditional acoustic representations, such as Fbank, or from large pre-trained acoustic models like WavLM, Whisper, HuBERT, and Wav2Vec. The workflow for the monomodal system training on Fbank features is shown below, which is the baseline model in this study.

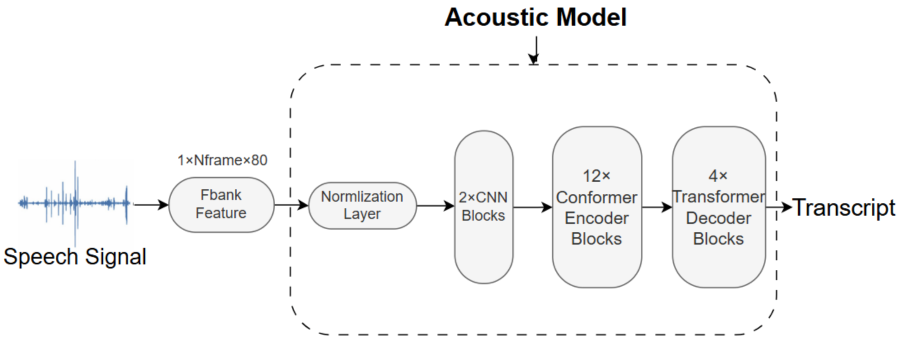


Figure 4.2: Pipeline of the Monomodal Model Trained on Fbank Features

4.3.2. MODEL TRAINING ON LARGE PRE-TRAINED MODEL EXTRACTED FEATURES

For the features extracted by large pre-trained models, the feature dimensionality is quite high, such as 1280 or 1024. To prevent the number of trainable parameters from becoming too large, an MLP projection layer is added to reduce the dimensionality to 80, which is the same dimension as the Fbank features. The pipeline is illustrated below.

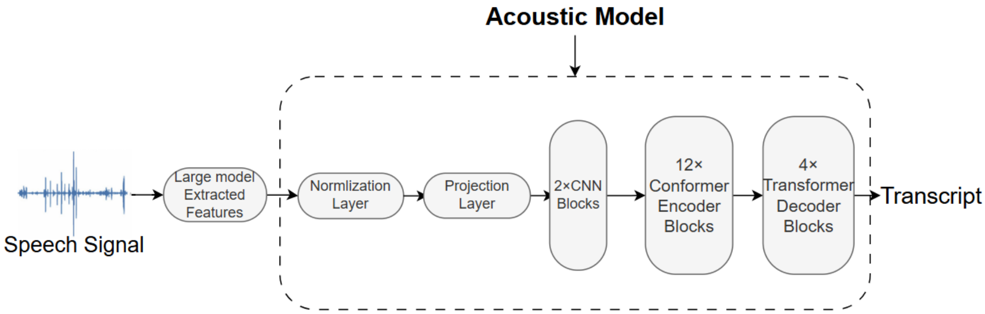

Acoustic Model
Large model Extracted Features
Speech Signal
Normlization Layer
Projection Layer
2×CNN Blocks
12× Conformer Encoder Blocks
4× Transformer Decoder Blocks
Transcript

Figure 4.3: Pipeline of the Monomodal Model Trained on Features Extracted by Pre-trained Model

4.3.3. MODEL FUSING WITH ARTICULATORY FEATURES

The multimodal approach fuses articulatory features derived from Electromagnetic Articulography (EMA) data with acoustic features. The workflow for the multimodal system is shown in the following figure. Three fusion methods are explored: the first method concatenates the acoustic and articulatory features before feeding them into the Conformer encoder, the second method uses a cross-attention mechanism to allow the model to attend to relevant information from both modalities dynamically, and the third method is using bidirectional cross-attention fusion strategy, as introduced in the previous chapter.

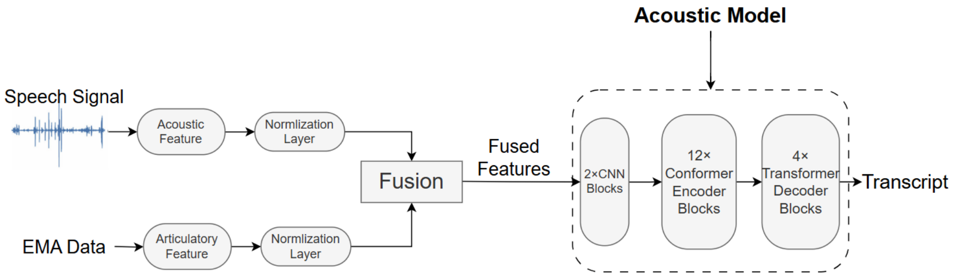


Figure 4.4: Pipeline of the Multimodal Model

4.3.4. Downstream Classification Task Using CNN

In this work, a downstream binary classification task was also designed to evaluate the separability of speech embeddings generated by the Conformer encoder. In this task, all parameters were frozen after the models were fully trained. Then the final layer output of the Conformer encoder served as input features, and the datasets used for this task corresponded directly to those of the primary experiment. Finally, these embeddings were used to train and test a simple CNN classifier to obtain the classification results. The pipeline of the downstream tasks is shown below.

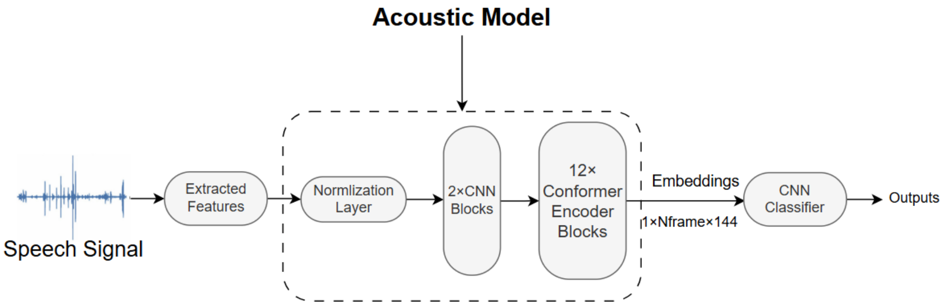


Figure 4.5: Downstream Classification Task Pipeline

4.4. Experiments setup

4.4.1. Dataset Split

In the TORGO dataset, the utterances were filtered to include only those with both EMA and audio data, resulting in a subset of 12,125 utterances. Additionally, for certain utterances, two types of microphones were used simultaneously: a head-mounted microphone and an array microphone. It means the audio transcription recorded with the two microphones is the same, only captured by different devices. If they appear separately in the training and test datasets, it can lead to a data leakage issue where the model is exposed to similar training data during testing, which will affect the fairness of the model evaluation. Therefore, these repeated audio recordings should be placed in the same dataset (e.g., only in the training or test set) to ensure data independence and evaluation reliability.

To ensure a balanced distribution of data across training, validation, and testing sets, the utterances in the TORGO dataset were split by speaker, with each severity level represented according to a 4:1:1 ratio. Specifically, each speaker's utterances were assigned to the training, validation, and testing sets in this proportion, maintaining an equal representation of severity levels. Also, for utterances recorded simultaneously by two types of microphones, the same utterance (from both microphones) was not split between the training and testing sets, thereby preventing data leakage and preserving the integrity of model evaluation. The following tables show the distribution of speakers of the training, validation, and testing sets.

Speaker	Severity Level	Gender	Training Set	Validation Set	Test Set
M01	Severe	Male	120	30	32
M02	Severe	Male	210	53	55
M04	Severe	Male	360	90	90
M05	M/S	Male	379	91	91
F03	Moderate	Female	468	116	117
F04	Mild	Female	325	80	82
M03	Mild	Male	530	132	136
FC01	Typical	Female	196	48	48
FC02	Typical	Female	1282	321	321
FC03	Typical	Female	997	251	251
MC01	Typical	Male	958	240	240
MC02	Typical	Male	740	189	189
MC03	Typical	Male	756	190	190
MC04	Typical	Male	825	206	206

Table 4.3: Distribution of utterances across training, validation, and test sets

4.4.2. TRAINING STRATEGY

An early stopping strategy was employed for training each model. At the end of each epoch, the model computes the WER on the validation dataset. If the WER does not improve for 5 consecutive epochs, the training will stop. Additionally, the training process will also be terminated if it reaches the maximum limit of 100 epochs. Each model is trained with a batch size of 8, using a learning rate of 1e-4 adjusted by a Noam scheduler that includes a warm-up phase of 25,000 steps. The loss function combines CTC loss with a weight of 0.3 and KL-divergence loss for label smoothing. The model uses Adam as an optimizer with betas set to (0.9, 0.98), and a weight decay of 1e-5 is employed to maintain stable updates throughout training.

4.4.3. EVALUATION METRICS

To evaluate model performance, the following evaluation metrics were employed:

- **Word Error Rate (WER):** WER is the primary metric used to assess the accuracy of

ASR systems. WER is calculated as follows:

$$\text{WER} = \frac{S+D+I}{N}$$

- S is the number of substitution errors,
- D is the number of deletion errors,
- I is the number of insertion errors,
- N is the total number of words in the reference transcription.

WER is the most widely used measure of recognition errors. In this study, it was used to compare the accuracy of ASR models.

- **Statistical significance:** In this work, significance testing was applied to determine whether differences in WER between different configurations (e.g., monomodal vs. multimodal models) are statistically significant. Statistical significance refers to whether the observed differences are likely due to chance or represent a real difference in performance.

  To compute p-values, the open-source tool WER-SigTest[45] was utilized, which provides a statistical testing script specifically for WER comparisons. The toolkit can calculate the p-value by computing the MAPSSWE[12](Matched-Pair Sentence Segment Word Error) between the results of the two ASR systems. The p-value calculation involves the following steps:

  1. **Calculate the WER difference for each sentence segment**:

  $$\Delta_i = \text{WER}_{1i} - \text{WER}_{2i}$$

  where i denotes each sentence segment.

  2. **Compute the mean of these differences**:

  $$\bar{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\Delta_i$$

  where N denotes the total number of segments.

  3. **Determine the standard deviation of the differences**:

  $$\sigma_\Delta = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\Delta_i - \bar{\Delta})^2}$$

  4. **Calculate the t-statistic**:

  $$t = \frac{\bar{\Delta}}{\frac{\sigma_\Delta}{\sqrt{N}}}$$

  The p-value is then obtained using this t-statistic to evaluate the significance of the performance difference. According to the criteria, significance levels are defined as follows: a p-value of 0.05 indicates significance at the 5% level, 0.01 indicates greater significance at the 1% level, and 0.001 denotes high significance at the 0.1% level.

5

Results and Discussion

This chapter analyzes the performance of Fbank features and features extracted from pre-trained models. It further examines their performance after incorporating articulatory features, with a detailed analysis of the impact of articulatory features on WavLM-extracted features. Finally, the chapter provides a comparison of three fusion methods, along with an in-depth analysis of their effectiveness.

5.1. Acoustic-only Models

The following table provides the WER results for all the acoustic-only models, including the baseline model trained on Fbank features and the models trained on features extracted by pre-trained models.

Input features	Severe (%)	M/S (%)	Moderate (%)	Mild (%)	Dys (%)	Typ (%)
Fbank-only	67.74	67.13	32.11	32.11	45.27	25.24
Whisper-only	56.51	**26.99**	15.14	7.52	27.36	16.55
WavLM-only	52.10	33.91	12.79	6.91	26.51	15.50
Wav2vec53	55.11	52.25	27.15	12.80	35.65	19.61
Hubert-only	**42.69**	29.41	**10.70**	**6.50**	**22.30**	**13.89**

Table 5.1: WER results of all acoustic-only models

A bar chart figure is also provided below to illustrate the result better. As shown in the figure, large pre-trained model extracted features outperform Fbank features for all the severity levels. This suggests that features extracted by large pre-trained models have advantages in capturing the complexity of the speech signal and lowering WER for dysarthric speech.

In the comparisons between features extracted by different pre-trained models, Hubert extracted features perform well in most dysarthric levels, especially the Severe level, but do not dominate all severity groups. For example, in the M/S, Moderate, and Mild groups, WavLM and Whisper have similar performances with Hubert, and Whisper even outperforms Hubert-only in the M/S group, suggesting that while Hubert-only is stronger overall, the other pre-trained models are still competitive in some specific severity groups. It's worth noting that the overall performance of Wav2vec2.0 features is relatively poor among all the pre-trained models, especially in the M/S and Moderate, and Mild groups, where its WER is higher than other large pre-trained models. This indicates that features extracted by Wav2vec2.0 are ineffective in dealing with dysarthric speech data.

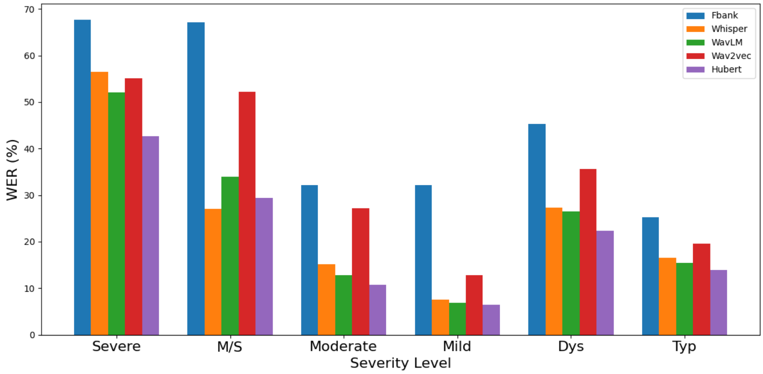
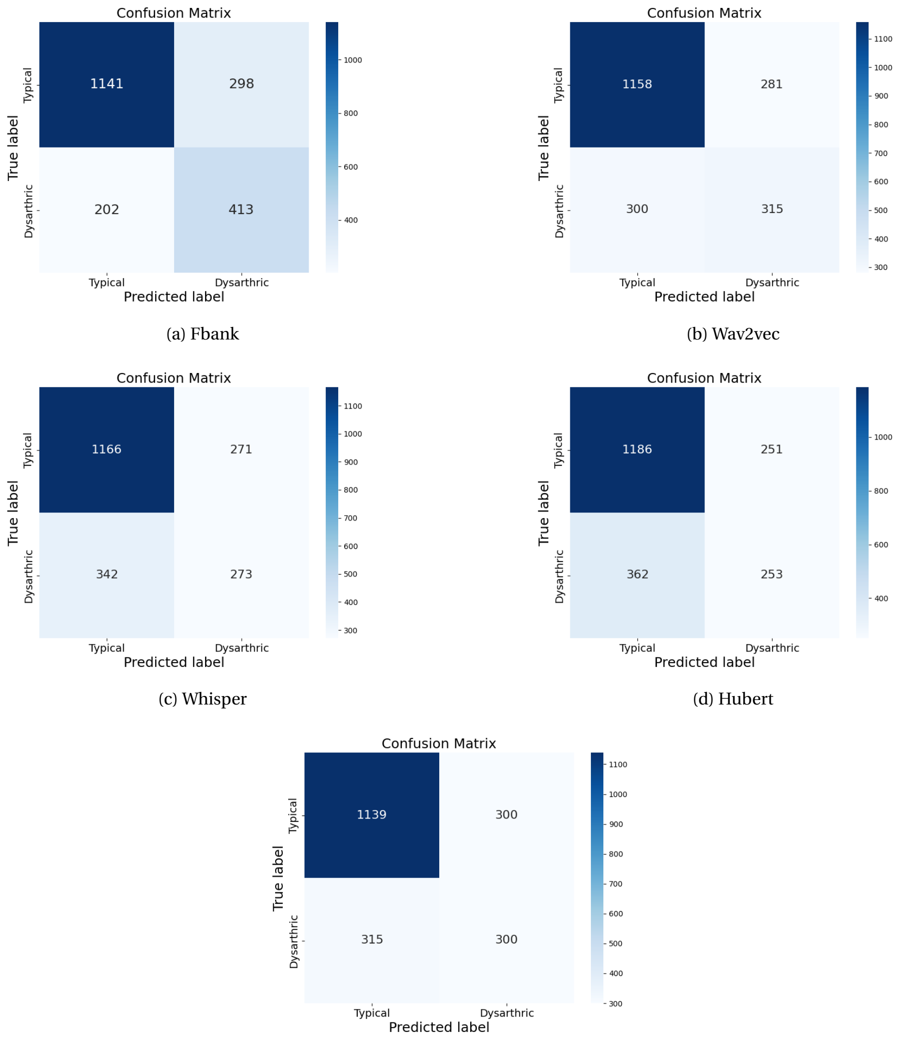


Figure 5.1: WER Comparison Across Different Models

For these acoustic-only models, corresponding downstream classification tasks were conducted. For each acoustic feature, embeddings encoded by Conformer encoders were used to train and test a CNN classifier, utilizing the same training, validation, and test utterances as the primary experiments. The confusion matrix is shown below. A notable observation is in the bottom-left corner of the confusion matrix, which denotes the number of dysarthric embeddings that are misclassified as typical ones. As shown in the figure, the Fbank feature has the lowest number of misclassifications, with only 202 cases, much fewer than those of the pre-trained model-extracted features. This indicates that, for Fbank features, the embeddings of typical speech are more distinct from those of dysarthric speech, making it easier for the classifier to distinguish between them. In contrast, embeddings derived from pre-trained model features show more similarity between typical and dysarthric speech, making it difficult for the classifier to categorize. This suggests that Fbank features are simple and retain big feature differences between typical and dysarthric after training; whereas features extracted by pre-trained models have higher generalization power, allowing them to capture more common features between typical and dysarthric speech, blurring the boundaries between them.



(a) Fbank

(b) Wav2vec

(c) Whisper

(d) Hubert

(e) WavLM

Figure 5.2: : Comparing Dysarthric Speech Classification Among Different Acoustic Feature Embeddings

5.2. EXPLORATION ON ARTICULATORY FEATURES

Before experiments on multimodal features, this work explored various combinations of articulatory features using data from five recording sensors on the upper lip, lower lip, tongue tip, tongue middle, and tongue back, which are directly involved in speech production. Specifically, this work selected the (x, y, z) coordinates and pairwise Euclidean distances for both the lip and tongue regions, resulting in a total of four feature combinations. The following table provides the results for these articulatory features combined with Fbank features. As shown in the table, when Fbank features are combined with articulatory features based on pairwise distances of sensors in lip regions, the model has the best performance. This result suggests that the distances between lip sensors capture more effective movement information, which is particularly helpful for speech recognition tasks. Based on this observation, this study will choose pairwise distances of sensors in lip regions as the articulatory features in the following experiments.

Input Features	**Fusion Strategy**	**Severe (%)**	**M/S (%)**	**Moderate (%)**	**Mild (%)**	**Dys (%)**	**Typ (%)**
Fbank+lip_dis	concat	**71.74**	82.01	**26.63**	20.93	**48.10**	**26.54**
Fbank+tongue_dis	concat	76.15	**66.78**	37.08	**20.12**	48.95	29.32
Fbank+lip_xyz	concat	75.55	82.70	38.64	27.24	54.00	29.94
Fbank+tongue_xyz	concat	77.96	76.82	66.84	22.15	58.68	29.35

Table 5.2: WER results of different combinations of articulatory features

5.3. Acoustic-articulatory Models

This section will show and analyze the results of the multimodal model to explore the effect of articulatory information on speech recognition performance to answer the first research question. The features in these multimodal models are all directly concatenated. The WER results together with acoustic-only models are shown in the following table. Table 5.2 shows that fusing articulatory features with acoustic features affects the WER results, but the effect varies for different models. For pre-trained models such as Wav2vec and Whisper, incorporating lip features had an inconsistent impact across different severity groups, sometimes leading to improvements, while at other times resulting in minimal change or even a slight decline. Also, it is worth noting that the Hubert-only monomodal model performs very well in several groups, especially in the Severe, Moderate, and Dysarthric groups, reaching 42.69%, 10.70%, and 22.30%, respectively. However, adding lip features (Hubert+lip_dis) did not noticeably improve performance and even slightly decreased it in some cases. This suggests that the Hubert model itself may be powerful enough for feature extraction of dysarthric speech, with less need for additional articulatory information.

Input features	Fusion strategy	Severe (%)	M/S (%)	Moderate (%)	Mild (%)	Dys (%)	Typ (%)
Fbank-only	-	67.74	67.13	32.11	32.11	45.27	25.24
Fbank+lip_dis	concat	71.74	82.01	26.63	20.93	48.10	26.54
Whisper-only	-	56.51	26.99	15.14	7.52	27.36	16.55
Whisper+lip_dis	concat	43.29	**17.99**	19.06	5.89	**22.24**	16.39
Wav2vec-only	-	55.11	52.25	27.15	12.80	35.65	19.61
Wav2vec+lip_dis	concat	49.10	61.59	21.41	10.98	33.61	22.67
WavLM-only	-	52.10	33.91	12.79	6.91	26.51	15.50
WavLM+lip_dis	concat	51.70	19.72	**10.70**	5.69	23.09	16.79
Hubert-only	-	**42.69**	29.41	**10.70**	6.50	22.30	**13.89**
Hubert+lip_dis	concat	50.70	28.03	11.49	**5.49**	24.35	14.11

Table 5.3: WER results for both acoustic-only and multimodal models using different fusion strategies.

Since WER results after adding lip features were quite mixed, it's necessary to calculate the p-value to see whether the observed differences were due to randomness or statistically significant. As shown in the table, only the articulatory features combined with WavLM extracted features have a p-value of 0.047, indicating a statistically significant improvement when lip distance features are added, leading to a lower WER. For features extracted by other models, the p-values exceed the 0.05 significance threshold, indicating that the performance variations are statistically insignificant. This implies that the observed changes are likely due to randomness rather than the influence of the added articulatory features. It's worth noting that, although adding articulatory features improves dysarthric speech most for Whisper extracted features, the p-value did not pass the significance threshold. This is because of the inconsistent changes in different severity groups which may affect the model's overall performance. Some extreme changes in certain samples influenced the p-value, making the overall performance appear improved, but not noticeably across all datasets. Due to the p-value results, this study will then focus on a detailed analysis of the WavLM model's feature performance with articulatory features. Since the p-value for the WavLM features is below the significance threshold, it is meaningful to specifically analyze in the next section how the articulatory features impact the acoustic features extracted by this model.

Features	p-value
Fbank+lip_dis	0.208
Whisper+lip_dis	0.066
Wav2vec+lip_dis	0.103
WavLM+lip_dis	**0.047***
hubert+lip_dis	0.697

Table 5.4: P-values for the comparison between acoustic-only and multimodal models on dysarthric speech, where * indicates p-values below the significance threshold, suggesting improved performance after fusing articulatory features.

5.4. IMPACT OF ARTICULATORY FEATURES ON WAVLM-EXTRACTED ACOUSTIC FEATURES

This section will analyze how articulatory features affect the WavLM extracted features. First, the test dataset embeddings were extracted, which had been encoded by the Conformer encoders' final layer. Then t-SNE was employed to visualize the embeddings of both acoustic-only and multimodal features. Different legends were used for speakers from different severity levels and genders, as shown in the following figures.

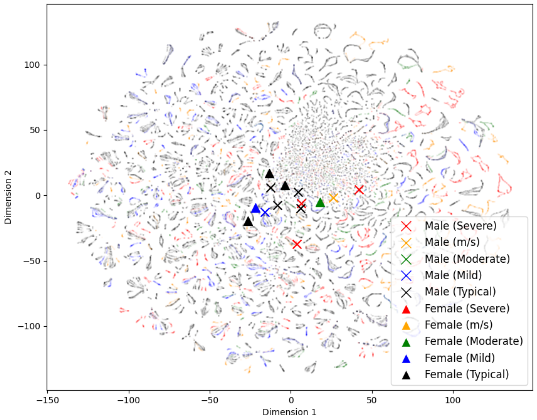


Figure 5.3: t-SNE Visualization of WavLM Acoustic-Only Embeddings

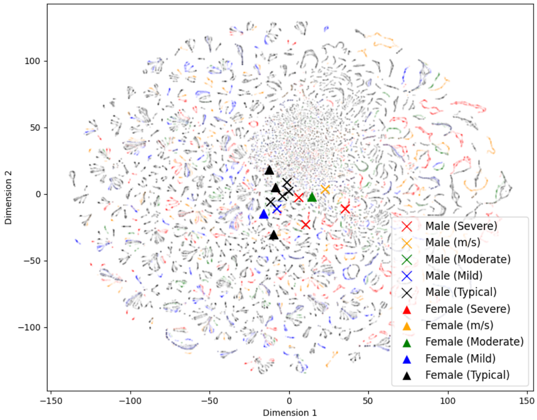


Figure 5.4: t-SNE Visualization of WavLM Embeddings Enhanced with Articulatory Features

As shown in the above figures, all the severity groups are represented in different colors. For each severity group, the centroid was computed, along with the distance between that centroid and the centroid of the typical group. The relative distance and WER reduction achieved by adding articulatory features are below. As shown in the figures below, the distances between typical speech and dysarthric speech decreased for all severity groups. This indicates that the articulatory information helped reduce the gap between the dysarthric and typical speech representations. Specifically, the Moderate group shows the highest relative reduction in feature distances (15.11%), followed by the Mild group (11.78%) and the M/S group (11.62%). It is worth noting that for speakers in the severe group, the distance has the smallest reduction(0.94%), which corresponds to the limited improvements observed in the WER results for severe-level speakers. The small reduction for the severe group suggests that the articulatory features alone may not be sufficient to bridge the large gap between severe dysarthric and typical speech.

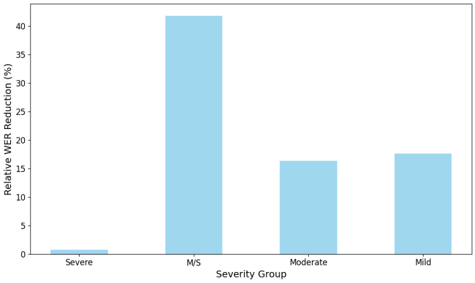


Figure 5.5: Relative Reduction in Centroid Distances Between Typical and Severity Groups for WavLM with Lip Features

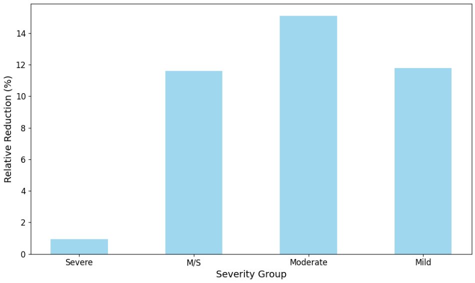
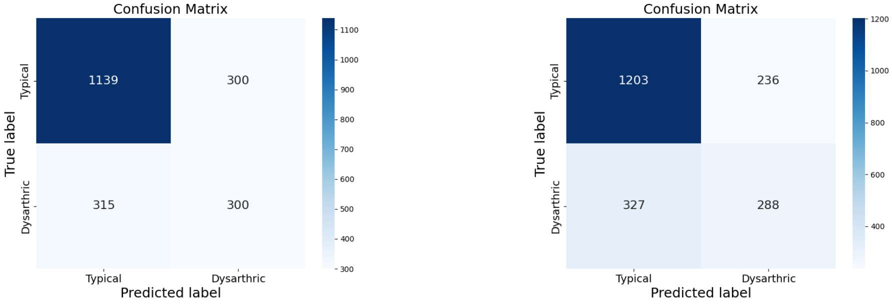


Figure 5.6: Relative WER Reduction for WavLM with Lip Features Across Severity Groups

This study also examined changes in centroid distances across gender groups to analyze the results from a gender perspective. However, the relative distance changes across gender groups are very minimal and do not show a clear pattern, suggesting that the severity level plays a more important role in the impact of articulatory features. Additionally, the unequal number of male and female utterances across severity groups makes it challenging to analyze the result solely from the perspective of gender groups.

In addition to t-SNE analysis, a downstream classification task was conducted. A simple CNN classifier was trained and tested using both multimodal and acoustic-only embeddings to classify speech embeddings as dysarthric or typical. As shown in the following figures, the CNN classifier training by multimodal embeddings can recognize more dysarthric utterances as typical ones, indicating that adding articulatory features improves the model's ability to deal with dysarthric speech, bringing these dysarthric samples closer to the distribution of typical speech, which is beneficial for the speech recognition tasks.



(a) Confusion Matrix of Acoustic-only Embeddings

(b) Confusion Matrix of Multimodal Embeddings

Figure 5.7: Comparing Dysarthric Speech Classification Using Multimodal vs. Acoustic-only Embeddings

5.5. Exploration on Fusion Methods

To answer the second research question, this work compared three fusion methods. The WER results are in the following table. As shown in the table, concatenation has the best performance on dysarthric speech at 23.09%, followed by bidirectional cross-attention(52.74%), and cross-attention(58.80%), indicating that simple concatenating can preserve the original information of multimodal features during the training process. It's worth noting that two cross-attention fusion strategies didn't perform well, suggesting that the correlation between WavLM extracted features and articulatory features is so weak that cross-attention may have difficulty learning meaningful interactions when computing attention weights. It could also be because cross-attention requires a large amount of data to learn complex relationships between features, and insufficient data may result in the distribution of attention weights being spread too thinly. Another possible reason is that the original articulatory feature has only 3 dimensions for each frame, while the acoustic feature has 80 dimensions. When the articulatory feature is projected to 80 dimensions using an MLP layer during cross-attention fusion, it may introduce some irrelevant information or noise, preventing the model from learning a valid alignment. These factors could contribute to the observed performance degradation. Among the two cross-attention methods, bidirectional cross-attention (DCA) outperforms traditional cross-attention (CA). This improvement is likely because DCA uses both articulatory features and acoustic features as query matrix, compensating for the lack of traditional CA which focuses only on the primary modality, ensuring that information from both modalities is fully utilized.

Input Features	Fusion Strategy	Severe (%)	M/S (%)	Moderate (%)	Mild (%)	Dys (%)	Typ (%)
WavLM+lip_dis	concat	**51.70**	**19.72**	**10.70**	**5.69**	**23.09**	**16.79**
WavLM+lip_dis	cross-attention	92.99	105.88	39.43	11.59	58.80	25.75
WavLM+lip_dis	bidirectional cross-attention	87.37	73.01	43.86	12.60	52.74	31.10

Table 5.5: Fusion strategy comparison for WavLM extracted features

6

Conclusion

In this research, a series of experiments are conducted to investigate the effectiveness of articulatory features when combined with pre-trained model-extracted acoustic features for dysarthric speech recognition tasks. First, I compared the performance of acoustic-only features, specifically Fbank features versus pre-trained model-extracted features. Then, I investigated the impact of different combinations of articulatory features and found that the distance between lip sensors had the best performance. Therefore, this feature was selected for use in subsequent experiments. Next, I evaluated the impact of combining articulatory features with each pre-trained model-extracted feature. Finally, I explored the effectiveness of different fusion strategies to integrate these features. These experiments are designed to answer the following research questions:

- **RQ1**: How effective are articulatory features in enhancing dysarthric speech recognition when combined with features extracted from pre-trained models?
- **RQ2**: How does the effectiveness of combining articulatory features with acoustic features vary across different severity levels of dysarthria?
- **RQ3**: What fusion methods can better utilize the feature information from both acoustic and articulatory features?

For **RQ1**, it's safe to conclude that articulatory features have the best performance on WavLM-extracted features, while for other pre-trained extracted features, articulatory features didn't take effect. The t-SNE analysis further supports this finding, demonstrating that articulatory features help reduce the gap between dysarthric and typical speech for WavLM-extracted features. For **RQ2**, when fused with WavLM extracted features, the results show that the articulatory features have better performance for M/S, Moderate, and Mild severity level, while the improvement for the severe level remains minimal. For **RQ3**, the results suggest that direct concatenating performs better than cross-attention-based fusion strategies across different gender groups across severe levels, indicating that simple concatenating may be a more robust option and complex mechanisms may not be effective in all cases.

For future work, I recommend exploring a broader range of combinations of articulatory features because I focused solely on evaluating the distances between sensors in the lip region in this study. The results indicate that the choice and combination of articulatory features significantly impact speech recognition performance. Therefore, investigating additional combinations of articulatory features would be a valuable direction for further research. Additionally, I recommend exploring a wider variety of models. In this study, I used the Conformer as the primary model. The choice of model can significantly influence the effectiveness of articulatory features. For instance, when traditional acoustic features were combined with articulatory features in prior work [55] which employed a different model, the articulatory features showed a positive impact. However, this effect was not observed in the case of the Conformer, suggesting that the model architecture plays a crucial role in the effectiveness of articulatory features. I also suggest collecting more data to address current limitations. A larger and more diverse dataset, including variations in gender, age, and severity levels, could also help analyze the effectiveness of articulatory features from different aspects.

Bibliography

[1] Adeleh Asemi et al. "Adaptive Neuro-fuzzy Inference System for Evaluating Dysarthric Automatic Speech Recognition (ASR) Systems: A case study on MVML based ASR". In: *Soft Computing* 2018 (May 2019). DOI: 10.1007/s00500-018-3013-4.

[2] Massa Baali et al. *Arabic Dysarthric Speech Recognition Using Adversarial and Signal-Based Augmentation*. 2023. arXiv: 2306.04368 [cs.SD]. URL: https://arxiv.org/abs/2306.04368.

[3] Leonardo Badino et al. "Integrating articulatory data in deep neural network-based acoustic modeling". In: *Computer Speech Language* 36 (2016), pp. 173–195. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2015.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0885230815000558.

[4] Alexei Baevski et al. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL]. URL: https://arxiv.org/abs/2006.11477.

[5] Mehala Balamurali and Arman Melkumyan. "t-SNE Based Visualisation and Clustering of Geological Domain". In: vol. 9950. Oct. 2016, pp. 565–572. ISBN: 978-3-319-46680-4. DOI: 10.1007/978-3-319-46681-1_67.

[6] Sanyuan Chen et al. "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (Oct. 2022), pp. 1505–1518. ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3188113. URL: http://dx.doi.org/10.1109/JSTSP.2022.3188113.

[7] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL]. URL: https://arxiv.org/abs/1406.1078.

[8] Shammur Absar Chowdhury, Nadir Durrani, and Ahmed Ali. "What do end-to-end speech models learn about speaker, language and channel information? A layer-wise and neuron-level analysis". In: *Computer Speech Language* 83 (2024), p. 101539. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2023.101539. URL: https://www.sciencedirect.com/science/article/pii/S088523082300058X.

[9] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

[10] Mengzhe Geng et al. *Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition*. 2022. arXiv: 2202.10290 [eess.AS]. URL: https://arxiv.org/abs/2202.10290.

[11] Yaroslav Getman et al. “Wav2vec2-based speech rating system for children with speech sound disorder”. In: *Interspeech*. International Speech Communication Association (ISCA). 2022, pp. 3618–3622.

[12] L. Gillick and S.J. Cox. “Some statistical issues in the comparison of speech recognition algorithms”. In: *International Conference on Acoustics, Speech, and Signal Processing,* 1989, 532–535 vol.1. DOI: 10.1109/ICASSP.1989.266481.

[13] Nada Gohider and Otman A. Basir. “Recent advancements in automatic disordered speech recognition: A survey paper”. In: *Natural Language Processing Journal* 9 (2024), p. 100110. ISSN: 2949-7191. DOI: https://doi.org/10.1016/j.nlp.2024.100110. URL: https://www.sciencedirect.com/science/article/pii/S294971912400058X.

[14] Anmol Gulati et al. *Conformer: Convolution-augmented Transformer for Speech Recognition.* 2020. arXiv: 2005.08100 [eess.AS]. URL: https://arxiv.org/abs/2005.08100.

[15] Pengcheng Guo et al. *The NPU-ASLP System for Audio-Visual Speech Recognition in MISP 2022 Challenge.* 2023. arXiv: 2303.06341 [eess.AS]. URL: https://arxiv.org/abs/2303.06341.

[16] Abner Hernandez et al. *Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition.* 2022. arXiv: 2204.01670 [cs.CL]. URL: https://arxiv.org/abs/2204.01670.

[17] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9 (Nov. 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[18] Harold Hotelling. “Analysis of a complex of statistical variables into principal components.” In: *Journal of Educational Psychology* 24 (1933), pp. 498–520. URL: https://api.semanticscholar.org/CorpusID:144828484.

[19] I-Ting Hsieh and Chung-Hsien Wu. “Dysarthric Speech Recognition Using Curriculum Learning and Articulatory Feature Embedding”. In: Sept. 2024, pp. 1300–1304. DOI: 10.21437/Interspeech.2024-444.

[20] Wei-Ning Hsu et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.* 2021. arXiv: 2106.07447 [cs.CL]. URL: https://arxiv.org/abs/2106.07447.

[21] Shujie Hu et al. *Exploring Self-supervised Pre-trained ASR Models For Dysarthric and Elderly Speech Recognition.* 2023. arXiv: 2302.14564 [cs.SD]. URL: https://arxiv.org/abs/2302.14564.

[22] Shujie Hu et al. *Self-supervised ASR Models and Features For Dysarthric and Elderly Speech Recognition.* 2024. arXiv: 2407.13782 [eess.AS]. URL: https://arxiv.org/abs/2407.13782.

[23] Yishan Jiao et al. *Simulating dysarthric speech for training data augmentation in clinical speech applications.* 2018. arXiv: 1804.10325 [eess.AS]. URL: https://arxiv.org/abs/1804.10325.

[24] Raymond D. Kent and Y J Kim. "Toward an acoustic typology of motor speech disorders". In: *Clinical Linguistics & Phonetics* 17 (2003), pp. 427–445. URL: https://api.semanticscholar.org/CorpusID:15842510.

[25] Tom Ko et al. "Audio augmentation for speech recognition". In: *Interspeech 2015*. 2015, pp. 3586–3589. DOI: 10.21437/Interspeech.2015-711.

[26] James A. Lowerre. "The Harpy Speech Recognition System". Ph.D. Dissertation. PhD thesis. Pittsburgh, PA: Carnegie Mellon University, 1976.

[27] Laurens van der Maaten and Geoffrey Hinton. "Viualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.

[28] Konstantin Markov, Jianwu Dang, and Satoshi Nakamura. "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework". In: *Speech Communication* 48.2 (2006), pp. 161–175. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2005.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0167639305001731.

[29] National Library of Medicine. *Dysarthria*. Accessed: 2025-02-15. n.d. URL: https://medlineplus.gov/ency/article/007470.htm.

[30] Claire Mitchell et al. "Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury". In: *Cochrane Database of Systematic Reviews* 1 (2017).

[31] Vikramjit Mitra, Hosung Nam, and Carol Espy-Wilson. "Robust speech recognition using articulatory gestures in a Dynamic Bayesian Network framework". In: *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings* (Dec. 2011). DOI: 10.1109/ASRU.2011.6163918.

[32] Vikramjit Mitra et al. "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 5205–5209.

[33] Laureano Moro-Velazquez et al. "Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson's Disease". In: *Interspeech 2019*. 2019, pp. 3875–3879. DOI: 10.21437/Interspeech.2019-2993.

[34] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356.

[35] Dhanesh Ramachandram and Graham W. Taylor. "Deep Multimodal Learning: A Survey on Recent Advances and Trends". In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108. DOI: 10.1109/MSP.2017.2738401.

[36] Siddharth Rathod, Monil Charola, and Hemant Patil. "Transfer Learning Using Whisper for Dysarthric Automatic Speech Recognition". In: Nov. 2023, pp. 579–589. ISBN: 978-3-031-48308-0. DOI: 10.1007/978-3-031-48309-7_46.

[37] Siddharth Rathod, Monil Charola, and Hemant A. Patil. “Noise Robust Whisper Features fornbsp;Dysarthric Severity-Level Classification”. In: *Pattern Recognition and Machine Intelligence: 10th International Conference, PReMI 2023, Kolkata, India, December 12–15, 2023, Proceedings*. Kolkata, India: Springer-Verlag, 2023, pp. 708–715. ISBN: 978-3-031-45169-0. DOI: 10.1007/978-3-031-45170-6_74. URL: https://doi.org/10.1007/978-3-031-45170-6_74.

[38] Frank Rudzicz, Aravind Namasivayam, and Talya Wolff. “The TORGO database of acoustic and articulatory speech from speakers with dysarthria”. In: *Language Resources and Evaluation* 46 (Jan. 2010), pp. 1–19. DOI: 10.1007/s10579-011-9145-0.

[39] Heidrun Schröter-Morasch and Wolfram Ziegler. “Rehabilitation of impaired speech function (dysarthria, dysglossia)”. In: *GMS current topics in otorhinolaryngology, head and neck surgery* 4 (Sept. 2005), Doc15.

[40] M. Schuster and K.K. Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093.

[41] University of Sheffield. *Articulatory Phonetics*. Accessed: 2025-02-15. n.d. URL: https://www.sheffield.ac.uk/linguistics/home/all-about-linguistics/about-website/branches-linguistics/phonetics/what-do-phoneticians-study/articulatory.

[42] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. “Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition”. In: *Interspeech 2018*. 2018, pp. 471–475. DOI: 10.21437/Interspeech.2018-1751.

[43] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[44] Tianzi Wang et al. *Hyper-parameter Adaptation of Conformer ASR Systems for Elderly and Dysarthric Speech Recognition*. 2023. arXiv: 2306.15265 [eess.AS]. URL: https://arxiv.org/abs/2306.15265.

[45] *WER statistical significance test*. https://github.com/talhanai/wer-sigtest. [online]. Accessed: 2024.

[46] Ronald J. Williams and David Zipser. “A Learning Algorithm for Continually Running Fully Recurrent Neural Networks”. In: *Neural Computation* 1.2 (1989), pp. 270–280. DOI: 10.1162/neco.1989.1.2.270.

[47] Alan Wrench and Korin Richmond. “Continuous speech recognition using articulatory data”. In: Oct. 2000, pp. 145–148. DOI: 10.21437/ICSLP.2000-772.

[48] Ji Wu. “Discriminative frequency filter banks learning with neural networks”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2019 (Jan. 2019). DOI: 10.1186/s13636-018-0144-6.

[49] Xurong Xie et al. “Variational Auto-Encoder Based Variability Encoding for Dysarthric Speech Recognition”. In: *Interspeech 2021*. interspeech$_2$021. ISCA, Aug. 2021, pp. 4808–4812. DOI: 10.21437/interspeech.2021-173. URL: http://dx.doi.org/10.21437/Interspeech.2021-173.

[50] Gaopeng Xu et al. “The NIO System for Audio-Visual Diarization and Recognition in MISP Challenge 2022”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–2. DOI: 10.1109/ICASSP49357.2023.10095577.

[51] Yufeng Yang, Peidong Wang, and DeLiang Wang. *A Conformer Based Acoustic Model for Robust Automatic Speech Recognition*. 2022. arXiv: 2203.00725 [cs.SD]. URL: https://arxiv.org/abs/2203.00725.

[52] Emre Yılmaz et al. *Articulatory Features for ASR of Pathological Speech*. 2018. arXiv: 1807.10948 [cs.CL]. URL: https://arxiv.org/abs/1807.10948.

[53] Chongchong Yu, Xiaosu Su, and Zhaopeng Qian. “Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition With Pre-Trained Models”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), pp. 1912–1921. DOI: 10.1109/TNSRE.2023.3262001.

[54] Zhengjun Yue et al. “Acoustic Modelling From Raw Source and Filter Components for Dysarthric Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2968–2980. DOI: 10.1109/TASLP.2022.3205766.

[55] Zhengjun Yue et al. “Multi-Modal Acoustic-Articulatory Feature Fusion For Dysarthric Speech Recognition”. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 7372–7376. DOI: 10.1109/ICASSP43922.2022.9746855.

[56] Jianxing Zhao et al. “A multiscale feature extraction algorithm for dysarthric speech recognition”. In: *Sheng wu yi xue gong cheng xue za zhi = Journal of biomedical engineering = Shengwu yixue gongchengxue zazhi* 40 (Feb. 2023), pp. 44–50. DOI: 10.7507/1001-5515.202205049.