Computer Speech and Language 95 (2026) 101839

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

# Raw acoustic-articulatory multimodal dysarthric speech recognition

Zhengjun Yue [a,*], Erfan Loweimi [b], Zoran Cvetkovic [c], Jon Barker [d], Heidi Christensen [d]

[a] Technology University of Delft, Delft, The Netherlands
[b] University of Edinburgh, Edinburgh, United Kingdom
[c] King's College London, London, United Kingdom
[d] University of Sheffield, Sheffield, United Kingdom

ARTICLE INFO

*Keywords:*
Multimodality
Acoustic-articulatory
Automatic dysarthric speech recognition
Acoustic modelling
Raw signal representations
Mutual information analysis

ABSTRACT

Automatic speech recognition (ASR) for dysarthric speech is challenging. The acoustic characteristics of dysarthric speech are highly variable and there are often fewer distinguishing cues between phonetic tokens. Multimodal ASR utilises the data from other modalities to facilitate the task when a single acoustic modality proves insufficient. Articulatory information, which encapsulates knowledge about the speech production process, may constitute such a complementary modality. Although multimodal acoustic-articulatory ASR has received increasing attention recently, incorporating real articulatory data is under-explored for dysarthric speech recognition. This paper investigates the effectiveness of multimodal acoustic modelling using real dysarthric speech articulatory information in combination with acoustic features, especially raw signal representations which are more informative than classic features, leading to learning representations tailored to dysarthric ASR. In particular, various raw acoustic-articulatory multimodal dysarthric speech recognition systems are developed and compared with similar systems with hand-crafted features. Furthermore, the difference between dysarthric and typical speech in terms of articulatory information is systematically analysed by using a statistical space distribution indicator called Maximum Articulator Motion Range (MAMR). Additionally, we used mutual information analysis to investigate the robustness and phonetic information content of the articulatory features, offering insights that support feature selection and the ASR results. Experimental results on the widely used TORGO dysarthric speech dataset show that combining the articulatory and raw acoustic features at the empirically found optimal fusion level achieves a notable performance gain, leading to up to 7.6% and 12.8% relative word error rate (WER) reduction for dysarthric and typical speech, respectively.

### 0.1. Introduction

Dysarthria is a speech disorder caused by neurological damage to the neuro-motor interface (Gowers, 2001). This neurological damage leads to the weakening or poor coordination of the muscles used for speaking (Duffy, 2013). Therefore, people with dysarthria often have diminished motor control over their speech articulators, producing speech with high variability that sounds markedly distinct from typical speech (Kent, 2000; Darley et al., 1969). Dysarthric speech is difficult to understand by machines

* Corresponding author.
*E-mail address:* z.yue@tudelft.nl (Z. Yue).

https://doi.org/10.1016/j.csl.2025.101839
Received 12 March 2025; Received in revised form 16 May 2025; Accepted 24 May 2025
Available online 10 June 2025

and people who are not familiar with a particular person's way of speaking. In addition, people with dysarthria often have physical disabilities such as those stemming from stroke or cerebral palsy. Consequently, they often have more restricted or involuntary body movements, limiting their ability to operate switches/handles and thus further impairing their interaction with machines. As a result, dysarthria constrains individuals in their social interactions with others and hinders their independence in daily life (Young and Mihailidis, 2010).

Automatic speech recognition (ASR) technology facilitates a voice-enabled interface that can be embedded into smart devices for hands-free human–machine interaction. This technology has the potential to greatly improve the quality of life for people with dysarthria by facilitating more effective communication with machines and other people. However, the current commercial ASR systems do not work reliably for dysarthric speech. Due to data scarcity, the significant dysarthric and typical speech mismatch and high speaker variability (Wilson and Blaney, 2000; Ferrier et al., 1995), automatic dysarthric speech recognition (ADSR) remains a significant challenge. The performance on dysarthric speech lags far behind the mainstream ASR systems for typical speech. As such there is an unmet need for reliable ASR systems capable of recognising dysarthric speech with high accuracy. Additionally, factors such as the intricate nature of dysarthric speech, varying degrees of speech motor control impairments, and the influence of co-occurring physical disabilities add to the complexity of ADSR.

There has been an expanding body of work aiming at improving dysarthric speech recognition performance. Various neural networks have been explored towards acoustic modelling for dysarthric speech in hybrid ASR (Espana-Bonet and Fonollosa, 2016; Joy and Umesh, 2018; Kim et al., 2018; Hermann and Magimai.-Doss, 2020; Yue et al., 2020a) and end-to-end (E2E) ASR systems (Almadhor et al., 2023; Shahamiri, 2021; Wang et al., 2023a; Hu et al., 2024; Shahamiri et al., 2023). In addition, a large amount of out-of-domain typical speech data has been exploited to effectively improve the recognition performance with more training data (Christensen et al., 2013; Yılmaz et al., 2019), and augmenting dysarthric speech data (Geng et al., 2022; Soleymanpour et al., 2022). Large-scale pre-trained ASR models, e.g., Wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023), have gained increasing attention to be adapted to improve the performance of dysarthric speech (Wang et al., 2023b; Vinotha et al., 2024).[1] Furthermore, various speech representations such as bottleneck features and features extracted from large-scale pre-trained models have also been demonstrated to be effective for dysarthric speech recognition in previous studies (Takashima et al., 2015; Yue et al., 2020a; Latha et al., 2023; Takashima et al., 2024; Patel and Patil, 2024).

Given that the acoustics of dysarthric speech are highly variable, the pronunciation of typical phonetic tokens often varies considerably. As a result, it is challenging to identify a reliable acoustic cue for a specific phone. Therefore, relying solely on the acoustic modality may not be a viable solution for acoustic modelling, especially with limited training data. To overcome this challenge and enhance model robustness, incorporating additional modalities alongside acoustic cues, is a compelling approach. The adoption of a multimodal framework allows for the integration of complementary channels of information, improving the model's capacity to discern and adapt to the intricate acoustic nuances inherent in dysarthric speech.

Attempts have been made to harness additional sources of knowledge in the speech production process. One such source of information is the articulatory information (Mitra, 2010), which captures the movements of speakers' articulators, e.g., lips and tongue. The positioning of the articulators plays an important role in human speech production and is strongly correlated with the linguistic content of speech. Compared with acoustic representations, articulatory space is comparatively simpler to model. The incorporation of articulatory space facilitates the modelling of the produced sounds, significantly enhancing the system's capability to recognise intricate speech patterns. Articulatory information has been shown to be more noise-robust (Wrench and Richmond, 2000), less speaker-variant (Fujimura, 1986) and more suitable to model the coarticulation variability (Kirchhoff et al., 2002; Frankel and King, 2001). Articulatory information therefore holds complementary information that can benefit ADSR systems.

Such multimodal speech recognition systems have received increasing attention over recent years owing to the fast development of deep learning techniques, which among others, provide effective frameworks for fusing different modalities. Previous studies have demonstrated the benefit of incorporating articulatory features by building acoustic-articulatory ASR systems for typical speech (Badino et al., 2016; Mitra et al., 2017). However, dysarthric speech recognition research has predominantly concentrated on applying acoustic representations alone.

The scarcity of parallel multimodal dysarthric data poses a significant obstacle to multimodal ADSR research. To mitigate the lack of *real* dysarthric articulatory recordings, previous studies have employed *synthetic* (pseudo) articulatory data derived from learnt acoustic-to-articulatory mappings, along with acoustic features for improving dysarthric speech acoustic modelling (Xiong et al., 2018; Yılmaz et al., 2018). By learning the mapping from acoustic to articulatory features, the synthesiser estimates articulatory data from the acoustic representations. In addition to the articulatory data synthesisers (such as *Gnuspeech* Hill et al., 2017 and TADA Nam et al., 2004), there have also been advanced neural architectures proposed for acoustic-articulatory inversion (Shahrebabaki et al., 2020, 2021; Udupa et al., 2023) which generate articulatory information. However, these synthetic articulatory features might not accurately capture the true articulatory characteristics of dysarthric speech. The mappings are usually learned using typical speech, leading to a significant mismatch when applied to dysarthric speech. In contrast, leveraging actual dysarthric articulatory data from individuals with dysarthria could provide a more precise and reliable representation.

[1] These models are trained on typical speech data with a single data modality. As this paper focuses on the integration of real articulatory data with raw acoustic features, our aim is to explore the synergistic effects of multimodal data tailored to dysarthric ASR rather than leveraging pre-trained mono-modal models optimised solely for general acoustic modelling. Additionally, comparing our results with those from pre-trained ASR models would introduce an out-of-domain performance perspective that falls outside the scope of this paper. By narrowing our focus, our paper emphasises the use of in-domain data, and there is no pre-trained model available for articulatory data. We did not apply or compare our results with models such as Whisper or Wav2vev2.0.

In the scope of dysarthric ASR, the integration of articulatory features with conventional hand-crafted acoustic features, such as MFCCs and Filterbanks (FBanks), has been a common practice (Xiong et al., 2018; Yue et al., 2022d). In this paper, we leverage the raw signal representations, including raw waveform with non-parametric and parametric CNNs (SincNet), raw magnitude spectrum, the raw source and filter components, and raw real and imaginary parts of Fourier Transform. These features avoid the loss of potentially relevant information that inadvertently occurs in traditional feature extraction methods, and potentially offer a comprehensive view of information encoded in the speech signal (Yousafzai et al., 2010; Ager et al., 2011). In particular, in the context of dysarthria, which introduces distortions and variations to speech, utilisation of the extra information offered by the raw signal representations, can be beneficial in capturing the nuances of the dysarthric speech and contributes to learning patterns which are more effective for ADSR. Acoustic modelling from raw signal representations has been proven effective in typical speech recognition (Sainath et al., 2015; Tüske et al., 2018; Loweimi et al., 2019, 2020b,a, 2021b), and also successfully applied in ADSR (Yue et al., 2022a; Loweimi et al., 2023a; Yue et al., 2022b). In this paper, we examine how these raw acoustic representations, when integrated with the real articulatory features, can improve the recognition of dysarthric speech in a multimodal setup.

We evaluate the performance of our multimodal dysarthric ASR system on the widely-used TORGO dataset (Rudzicz et al., 2012b). It includes aligned acoustic and articulatory data for both dysarthric and typical speech. There have been several TORGO-based studies demonstrating the usefulness of incorporating articulatory information alongside handcrafted acoustic features for the ADSR task (Rudzicz, 2010a,b; Rudzicz et al., 2012a).

In summary, the existing research gap in multimodal acoustic-articulatory acoustic modelling for ADSR has three primary aspects:

1. The need for models that capture dysarthric articulation. Previous work (Yılmaz et al., 2018; Xiong et al., 2018) modelled dysarthric articulation parameters from acoustic signals using knowledge from typical speech. This raises concerns about the authenticity of synthesised articulatory data in reflecting actual dysarthric speech properties.
2. The exploration of advanced and robust acoustic models incorporating real articulatory data. Although previous studies (Rudzicz, 2010a; Rudzicz et al., 2012a) showed benefits in incorporating the real articulatory data for the ADSR task, advanced modelling of this modality remains unexplored.
3. The optimal approach for fusing the acoustic and articulatory features has not been thoroughly explored in the context of ADSR.

The main contributions of this paper are summarised below:

- We systematically compare the articulatory movement pattern mismatch between dysarthric and typical speech. This is achieved through the visualisation of the 3D point cloud of several real articulatory data samples and an analysis of the statistical space distribution across the dataset, employing the Maximum Articulator Motion Range (MAMR) as a metric (Duan et al., 2020; Teplansky et al., 2020).
- We conduct comprehensive articulatory information analysis, including the computation of *mutual information*, to identify and select the most task-beneficial articulators, thereby enabling a more informed integration of articulatory features with acoustic features for optimal multimodal ADSR performance.
- Training dynamics of the models in terms of cross-entropy (CE) loss and word error rate (WER), with and without articulatory features, is illustrated for both dysarthric and typical speech.
- We investigate multimodal acoustic modelling for ADSR using real articulatory information in combination with acoustic features, encompassing both hand-crafted features and raw signal representations, e.g., raw waveform with non-parametric and parametric CNNs, magnitude spectrum, the raw source and filter components, and raw real and imaginary parts of the Fourier transform.

The rest of this paper is organised as follows. The real articulatory data in the TORGO dataset is analysed in Section 1 and the potential advantages when such information is applied to dysarthric speech recognition is discussed. Section 2 presents the architecture and the process of developing multimodal acoustic-articulatory ADSR systems. The experimental setup, results and discussion are presented in Section 3. The paper concludes with Section 4, summarising the key findings and proposing avenues for future work.

## 1. TORGO dataset

### 1.1. Data description

TORGO (Rudzicz et al., 2012b) is a widely used dysarthric speech dataset comprising both isolated words and sentence utterances. The recordings are collected from 15 speakers. Eight of the speakers (5 males, 3 females) suffer from dysarthria ranging from mild to severe, while others are typical speakers (4 males, 3 females). TORGO contains aligned acoustic and articulatory recordings for 80.2% of the utterances (13127 out of 16363). Articulatory recordings are not available for all of the speakers, and electromagnetic midsagittal articulography (EMA) (Schönle et al., 1987) data is missing for some sessions. The missing EMA data is usually removed due to the dropped sensors or the disturbing magnetic field during the recording session. Therefore, the 13127 utterances with both articulatory and acoustic data will be used in this work.

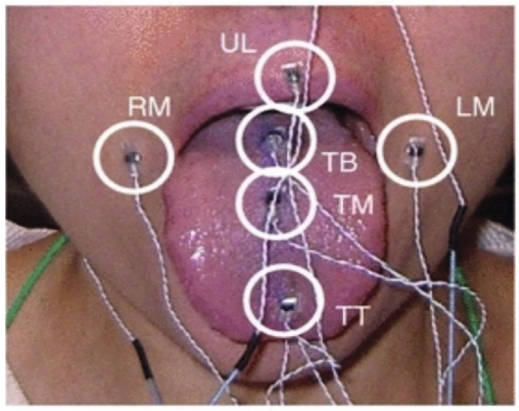



**Fig. 1.** The placement of coils on the RM, LM, UL, TT, TM and TB in the AG500 EMA system.
*Source:* Figure adapted from (Rudzicz et al., 2012b).

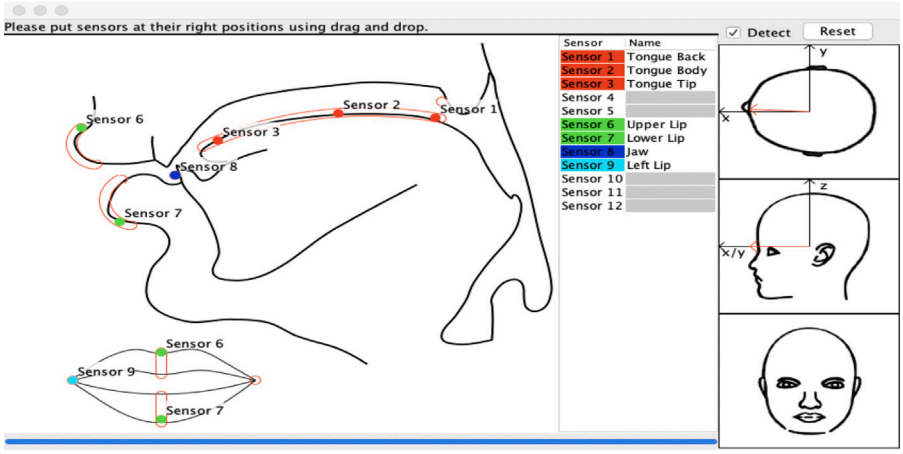



**Fig. 2.** Visualisation of sensors configuration using Visartico.

### 1.2. EMA data

The articulatory data in TORGO, referred to as EMA data, was collected through a 3D AG500 EMA system. The system captures the movement of 12 sensors placed on key articulators in the 3D space, each returning sensor positions in Cartesian coordinates (x, y, z) along with the spatial orientation angles. The sensors were attached to the tongue back (TB), tongue middle (TM), tongue tip (TT), forehead, bridge of the nose (BN), upper lip (UL), lower lip (LL), lower incisor (LI), left mouth (LM), right mouth (RM), left ear (LE) and right ear (RE). Fig. 1 illustrates the placements of several key sensors (RM, LM, UL, TT, TM and TB) attached to the articulators.

Before using the data, it is helpful to verify the accuracy of sensor labels and identify any malfunctioning sensors. Visartico (Ouni et al., 2012) is a useful articulatory data visualisation toolkit, and facilitates the confirmation of sensor placements and the detection of any faults. The sensor layout as configured within Visartico is presented in Fig. 2.

Upon detailed examination of the raw EMA data from each sensor channel, we made the following observations:

1. **Inconsistencies in sensor data ordering between typical and dysarthric group.**
   Table 1 presents the channel sequence for the dysarthric and typical groups. Notably, the 4th, 5th and 10th sensors are attached differently (highlighted in red for clarity). This observation necessitates caution when using EMA data across all speakers and sensors. However, the sensor sequence associated with tongue and lip articulators (specifically the 1st, 2nd, 3rd, 5th and 10th sensors) remains consistent between the two speaker groups. This mitigates concerns about mismatching the sensor order when only using lip or tongue sensor data.
2. **Issues with specific sensors.** In certain cases, sensors were found to shift unexpectedly during recordings, particularly for speakers M04 and MC01, where the TB sensor dislodged from the intended position. This could result from external electromagnetic signal interference or physical displacement of the sensor. Additionally, instances of non-functional sensors

**Table 1**
The EMA data channel sequence attached for the typical and dysarthric groups.

| Channel No. | Typical | Dysarthric |
|---|---|---|
| 1 | Tongue back (TB) | Tongue back (TB) |
| 2 | Tongue middle (TM) | Tongue middle (TM) |
| 3 | Tongue tip (TT) | Tongue tip (TT) |
| 4 | Right mouth (RM) | Forehead |
| 5 | Forehead | Bridge of the nose (BN) |
| 6 | Upper lip (UL) | Upper lip (UL) |
| 7 | Lower lip (LL) | Lower lip (LL) |
| 8 | Lower incisor (LI) | Lower incisor (LI) |
| 9 | Left mouth (LM) | Left mouth (LM) |
| 10 | Bridge of the nose (BN) | Right mouth (RM) |
| 11 | Left ear (LE) | Left ear (LE) |
| 12 | Right ear (RE) | Right ear (RE) |

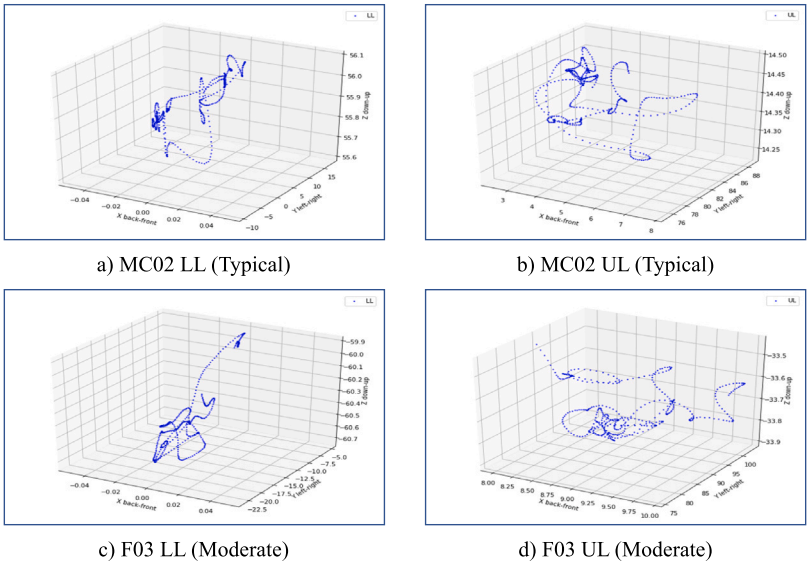



**Fig. 3.** 3D point cloud of the UL and LL for the utterance "The pair of shoes was new" for speakers MC02 and F03.

were noted, such as with speaker M05, where the TM and TB sensors failed to record any data, resulting in zero values for these channels.

3. **Reliability of lip sensor data.** The sensors placed on the upper lip (UL) and lower lip (LL) demonstrated consistent attachment across all speakers. This suggests that the lip articulatory data may be more reliable than the tongue data as articulatory features.

Sensors attached to the forehead, bridge of the nose (BN), left ear (LE), and right ear (RE) primarily serve to capture head movements and are used for reference purposes. Consequently, our forthcoming feature analysis will focus on articulatory features derived from the lip (UL and LL) and tongue (TT, TB, and TM) regions.

### *1.3. 3D point cloud*

To provide an intuitive insight into the articulation differences between dysarthric and typical speech within the TORGO dataset, we visualised the 3D point cloud of selected articulatory data samples. Fig. 3 depicts the 3D trajectories of the UL and LL sensors during the articulation of the phrase *"The pair of shoes was new"* by two speakers: F03 (moderate) and MC03 (typical).

It is seen that the typical speaker demonstrates clearer lip movements with less overlap in the 3D trajectory, indicating more precise and controlled articulation. Furthermore, the movement range of the typical speaker's lip articulators is notably constrained in the horizontal (left–right) and vertical (up–down) planes compared to that of the speakers with dysarthria. However, the movements along the $X$ (front–back) axis do not exhibit a marked difference between the two groups. These observations suggest that the typical speakers can produce speech with minimal mouth movements, whereas speakers with dysarthria may experience challenges in controlling mouth opening and closing during speech, leading to exaggerated and less coordinated movements.

The 3D point cloud plots provide an insightful preliminary tool for understanding the articulatory differences between dysarthric and typical speech. The visualised samples, although illustrative, do not capture the full variability and complexity inherent in

**Table 2**
The number of utterances where the prompts are overlapping between speaker MC02 and other speakers.

| Spk | M01 | F03 | F04 | M03 | FC02 | FC03 | MC01 |
|---|---|---|---|---|---|---|---|
| MC02 | 80 | 274 | 209 | 339 | 756 | 579 | 555 |

**Table 3**
The MAMR statistics ($\mu$ and $\sigma$ in mm) for different articulators for the dysarthric and typical speech.

| Articulator | Direction | Dysarthric speech | | Typical speech | |
|---|---|---|---|---|---|
| | | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| TT | X | 13.1 | 16.2 | 10.7 | 7.6 |
| | Y | 6.4 | 10.0 | 10.1 | 9.8 |
| | Z | 1.2 | 1.9 | 0.7 | 1.0 |
| TM | X | 6.1 | 11.5 | 5.7 | 4.5 |
| | Y | – | – | – | – |
| | Z | 1.7 | 3.7 | 0.7 | 2.6 |
| TB | X | 17.8 | 30.5 | 11.9 | 5.6 |
| | Y | 15.3 | 13.2 | 13.4 | 5.7 |
| | Z | 2.4 | 4.2 | 2.0 | 0.6 |
| UL | X | 6.3 | 11.5 | 7.4 | 5.5 |
| | Y | 21.1 | 26.3 | 18.7 | 8.7 |
| | Z | 0.8 | 2.6 | 0.5 | 2.9 |
| LL | X | – | – | – | – |
| | Y | 19.3 | 10.6 | 26.6 | 10.3 |
| | Z | 0.7 | 1.5 | 0.4 | 0.4 |

the articulatory movements of all dysarthric and typical speech utterances. Recognising this limitation, a more comprehensive quantitative analysis is necessary to accurately characterise and compare the articulatory motion patterns across the entire dataset. Such an analysis will provide a deeper understanding of the differences in speech articulation between speakers with dysarthria and typical speakers.

### *1.4. Statistical articulatory space distribution*

To conduct a quantitative analysis of the statistical articulatory space distribution in dysarthric and typical speech, particularly within the lip and tongue regions, we use the MAMR (Duan et al., 2020; Teplansky et al., 2020) as a key metric. MAMR quantitatively captures the articulatory dynamics by measuring the span between the maximal and minimal positions of an articulator within a single utterance. We will provide a brief explanation of the data preparation process and compare the MAMR statistics among the selected data samples from dysarthric and typical groups.

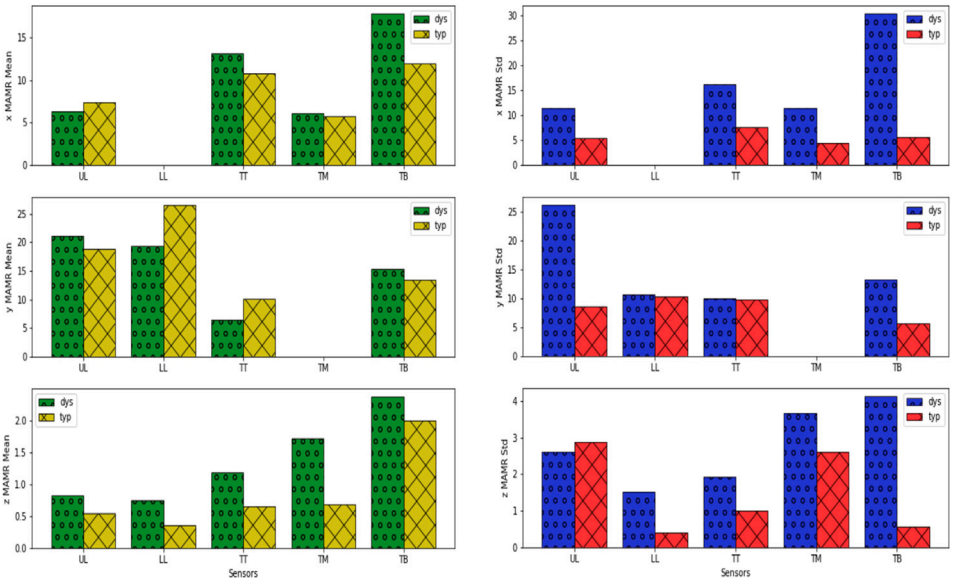
**Data Preparation:** To ensure a fair comparison, our analysis focuses on a subset of utterances that share the same prompts across different speakers. Due to issues with sensor reliability, speakers M04 and M05 were excluded from this subset. The final subset for this analysis includes recordings from four dysarthric speakers (M01, F03, F04, M03) and four typical speakers (MC01, MC02, FC02, FC03).

To standardise the comparison, we selected MC02, the speaker with the highest number of prompt-overlapping utterances with other speakers, as the reference for specifying utterances. Table 2 details the number of shared prompts between MC02 and the other speakers in the subset, facilitating a systematic analysis on these overlapping utterances. For instance, '80' indicates that there are 80 utterances where the prompts overlap between speakers M01 and MC02, providing a common ground for the comparative MAMR analysis.

***Maximum articulator motion range (MAMR)*:** We calculate the mean ($\mu$) and standard deviation (STD) ($\sigma$) for selected articulators (TT, TM, TB, UL and LL[2] along three spatial axes: $X$ (front–back), $Y$ (left–right) and $Z$ (up-down). The values are averaged for the typical speech (speaker MC02) and the dysarthric speech (speakers M01, F03, F04 and M03) and are presented in Table 3. Additionally, Figs. 4(a) and 4(b) visually depict these comparisons, illustrating the differences in articulatory motion patterns across the specified directions for both groups.

Distinct patterns in articulatory space between dysarthric and typical speech are observed by analysing MAMR. The standard deviations ($\sigma$) of MAMR values for dysarthric speech are generally larger than those of typical speech across most articulators and directions, with the exception of the UL sensor in the Z (up-down) direction. This higher $\sigma$ in dysarthric speech implies greater variability in articulatory movements, indicating the fluctuating nature of dysarthric articulation. A particularly pronounced difference is observed in the TB sensor along the $X$ (front–back) direction, where dysarthric speech exhibits a $\sigma$ of 30.5 mm,

[2] Among the 12 sensor coils detailed in Table 1, five (UL, LL, TT, TM, TB) are selected for articulatory features analysis while the remainder serve as reference sensors.




(a) MAMR $\mu$. (b) MAMR $\sigma$.

**Fig. 4.** MAMR statistics between dysarthric and typical speech.

significantly higher than the 5.6 mm observed in typical speech. This indicates that the speakers with dysarthria have less control over the posterior tongue muscles when producing speech.

When examining the mean ($\mu$) values, dysarthric speech consistently shows higher MAMR across all sensors in the $Z$ (up-down) direction, illustrating an expanded range of vertical articulator movements compared to typical speakers. This finding highlights the tendency of speakers with dysarthria to engage in more pronounced up-and-down articulator motions. However, in the $X$ (front–back) and $Y$ (left–right) directions, dysarthric speech does not uniformly exhibit higher MAMR values compared to typical speech, indicating a more complex relationship between articulatory motion patterns and speech types along these axes.

We also observe a narrower span of the motion range in the $Z$ direction for all five evaluated articulators, in contrast to the $X$ and $Y$ directions. For example, the TT sensor demonstrates mean MAMR values of 1.2 mm in the $Z$ direction, compared to 13.1 mm and 6.4 mm in the $X$ and $Y$ directions for dysarthric speech, and 0.7 mm in the $Z$ direction, against 10.7 mm and 10.1 mm in the $X$ and $Y$ directions for typical speech. This trend is corroborated by the standard deviation values, which also indicate reduced variability in the $Z$ direction for both speech groups.

Due to large speaker variability, we compare the statistics across the different severity levels. By comparing the MAMR distribution statistics for various speakers, we found distinct articulatory movement patterns at various severity levels. The variability in the mean values of all five articulators is much larger among dysarthric speakers, compared to typical speakers across all three spatial dimensions. Specifically, the difference in means between moderately and mildly dysarthric speech is smaller compared to the differences observed between severely and moderately, and between severely and mildly dysarthric speech. This pattern indicates the significant speaker variability inherent within dysarthric speech.

It is worth noting that while moderately and mildly dysarthric speech may share similarities with each other and, to some extent, with typical speech, severely dysarthric speech is characterised by substantially higher MAMR $\mu$ and $\sigma$ values. These findings indicate that speakers with severe dysarthria have less control over their articulators, leading to greater variability in their speech patterns.

The 3D visualisation of selected samples presented in Fig. 3 and the analysis of MAMR statistics reveal notable differences in the speech production from an articulatory perspective. These analyses not only highlight the constraints on articulatory movements among speakers with dysarthria but also illustrate the articulatory mismatch between dysarthric and typical speech. As a result, leveraging articulatory information has the potential to help towards distinguishing between the typical and dysarthric speech as well as capturing and normalising the dysarthric speech variability. These insights could, in turn, contribute valuable information for the speech recognition task, enhancing overall performance.

### *1.5. Articulatory information analysis*

To improve ADSR performance using articulatory features, it is essential to identify which articulators retain substantial phonetic information. We approached this analysis from two perspectives: (1) determining which articulators are less distorted when comparing dysarthric speech with typical speech, and (2) evaluating their mutual information (MI) (Kraskov et al., 2004; Ross, 2014; Pedregosa et al., 2011) with phonetic classes as a proxy for phonetic information content.

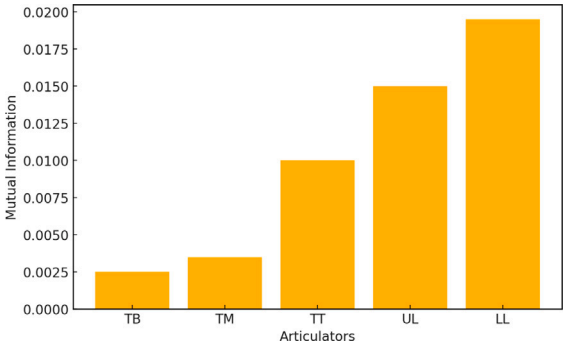



**Fig. 5.** MI between typical and dysarthric speech across different articulators.

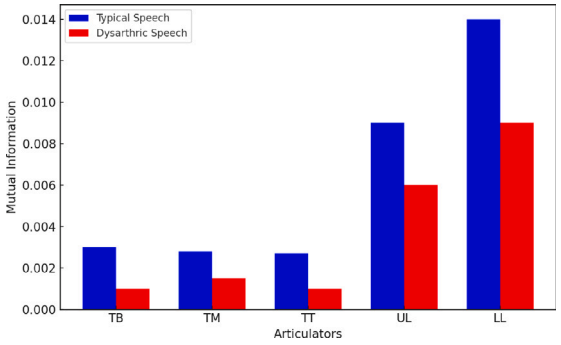



**Fig. 6.** MI between articulator positions and the phone alignment trained for ASR on typical and dysarthric speech.

### *1.5.1. MI analysis between typical and dysarthric speech*

MI between dysarthric and typical speech for different articulatory features serves as a powerful statistical metric that quantifies the robustness of these features: a higher MI value indicates that an articulatory feature remains closely aligned with typical speech, suggesting minimal distortion from dysarthria. As such this analysis helps determine which articulators are less affected by dysarthric conditions.

Fig. 5 illustrates the MI values for various articulators. LL and UL stand out with the highest MI values, suggesting that they are less affected by the distortions associated with dysarthric speech. This implies that UL and LL are more reliable for recognising dysarthric speech patterns, as they remain more stable between typical and dysarthric speech. Consequently, incorporating these lip-based articulators in an ADSR system is more beneficial.

### *1.5.2. MI with phone alignments*

To statistically explore how articulators contribute to phonetic distinctions, we computed the MI between articulatory features and phone alignments, separately for typical and dysarthric speech. This analysis highlights each articulator's contribution to phonetic distinction – a critical factor for ADSR performance.

Fig. 6 illustrates the MI values between various articulators and phone alignments (per frame). As anticipated, mutual information is higher for typical speech compared to dysarthric speech, reflecting the more organised and consistent articulatory movements in typical speakers. Furthermore, comparing the MI values among different articulators reveals a notable pattern: LL and UL demonstrate higher mutual information with phone alignments than tongue-based articulators (TB, TM and TT). This indicates a stronger association with phonetic information and a greater ability to represent phones accurately.

To summarise, lip-based articulators (UL and LL) are less distorted between typical and dysarthric groups and show a higher capability of representing phonetic information. Combined with the observed reliability of lip-based sensors discussed in Section 1.2, therefore, lip-based articulators should be a better choice in building ADSR systems. The ADSR results in Table 5 demonstrate how channel selection affects recognition performance and further support the overall observations.

Additionally, based on the above observations, we suggest that collecting detailed articulatory data may not be strictly necessary. Instead, video data capturing lip movements, which is easier and more practical to obtain, could effectively provide the critical articulatory information required for multimodal ADSR systems. This shift in data collection strategy could significantly reduce complexity while maintaining or even enhancing the system's performance by leveraging visually accessible articulatory features.

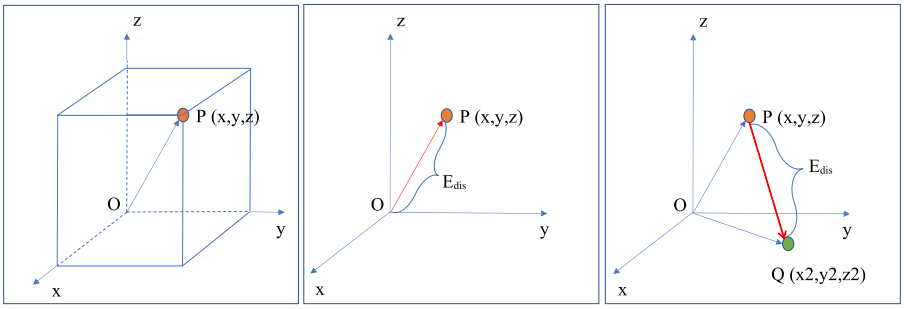



**Fig. 7.** Three measures of EMA data. $E_{dis}$: Euclidean distance.

## 2. Multimodal acoustic-articulatory ADSR

The preliminary observations discussed in the previous section suggest that the articulatory features may hold valuable complementary information for ADSR systems, especially lip-based articulators. To explore this potential and further validate the previous observations from MI analysis, we investigate the usefulness of articulatory information by building various multimodal acoustic-articulatory ASR systems tailored for dysarthric speech.

### 2.1. Data processing

The integration of raw EMA data into the ASR systems necessitates appropriate pre-processing to ensure its effectiveness as a feature. This pre-processing involves three steps: low-pass filtering, downsampling, and channel selection (Yue et al., 2022d). For channel selection, the EMA dataset comprises 12 channels, each corresponding to a specific sensor attached to an articulator. Considering the varied roles of different articulators in speech production, not all channels may contribute effectively to the recognition task. To avoid redundant information, we selectively incorporate those EMA channels that are most beneficial to the recognition task (supported by the MI analysis in Section 2.E and the ASR performance). The channel selection is performed in conjunction with the development of the acoustic-articulatory speech recognition systems, ensuring that only the most relevant articulatory information is utilised. Further details on the channel selection process and its impact on the recognition performance are presented in Section 3.2.

### 2.2. Optimal representations for the articulatory features

There are different ways to employ the processed EMA data. We consider the following as articulatory features:

1. The Cartesian coordinates (x,y,z) positions. (E.g., *Lip*)
2. The Euclidean distance between the articulatory sensors and the origin (0,0,0) (origin Euclidean distance). (E.g., *Lip_EuD_origin*)
3. The pair-wise Euclidean distance between sensors. (E.g., *Lip_EuD*)

Fig. 7 shows these three articulatory measures in the 3D space.

The Cartesian coordinates (either (x,y) or (x,y,z)) of the articulators are widely used as articulatory features in the previous ADSR studies (Rudzicz, 2009; Rudzicz et al., 2012a). Instead, in the following we use the Euclidean distance between the articulator and the origin and the pair-wise Euclidean distance between two articulators (e.g., in the lip region[3]).

Fig. 8 depicts the MAMR distribution of the three measures (top to bottom refers to the measurements from left to right in Fig. 7) of the lip sensors (UL and LL) along the front–back (X) direction for dysarthric and typical speech separately. The distributions follow the log-normal probability density function (Johnson et al., 1995). Comparing the envelopes of the MAMR distribution, the pair-wise Euclidean distance appears to be able to better distinguish between dysarthric and typical speech which implies that it can capture the mismatch between the two types of speech in the articulatory space. Although the MAMR distributions of the dysarthric and typical speech groups of the origin Euclidean distance are the most discriminative among the three measures, they are more overlapping than the pair-wise Euclidean distance. It is also observed from the middle graph in Fig. 8 that the MAMR distribution has a larger standard deviation for the dysarthric speech than the typical speech, and there tend to be more utterances with high MAMR (e.g., higher than 100 mm). This is owing to more abnormal MAMR values in the dysarthric speech.

Table 4 compares the MAMR mean ($\mu$) and STD ($\sigma$) of the three articulatory measures of the lip sensors (*Lip, Lip_EuD_origin, Lip_EuD*) for dysarthric and typical speech. The STD of the dysarthric speech is higher than the typical speech across the three

[3] The Euclidean distance between the UL (UL_x,UL_y,UL_z) and LL (LL_x,LL_y,LL_z) articulators.

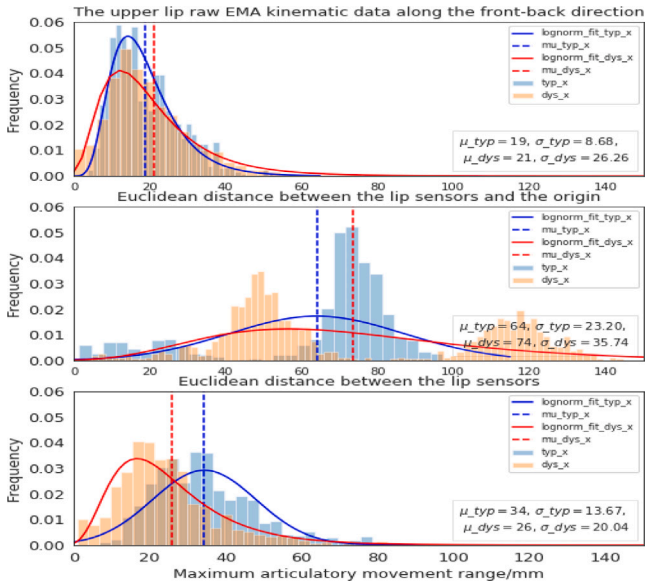



**Fig. 8.** MAMR distribution map of three articulatory measures.

**Table 4**
The MAMR statistics of the three articulatory measures of the lip sensors for dysarthric and typical speech.

| Articulator | Dysarthric speech | | Typical speech | |
|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Lip | 21 | 26.3 | 19 | 8.7 |
| Lip_EuD_origin | 74 | 35.7 | 64 | 23.2 |
| Lip_EuD | 26 | 20.0 | 34 | 13.7 |

measures. The higher STD implies that speakers with dysarthria exhibit more fluctuation than typical speakers. The MAMR mean of the Cartesian coordinates and the origin Euclidean distance for the dysarthric speech is larger than that for typical speech. In contrast, the MAMR mean of the pair-wise Euclidean distance for the dysarthric speech is smaller than that for typical speech. The MAMRs of Cartesian coordinates and the origin Euclidean distance measure the absolute articulators displacement, while the pair-wise Euclidean distance gauges the relative displacement.

This suggests that the speakers with dysarthria tend to move or shake their bodies or heads while speaking and should work harder to move their articulators, resulting in smaller MAMR means. This demonstrates the capacity of the pair-wise Euclidean distance feature to mitigate and implicitly normalise the effects of body and head movements.

### *2.3. Acoustic and articulatory feature integration*

Our multimodal acoustic model draws inspiration from the multi-stream acoustic modelling (Loweimi et al., 2020b, 2021a,b) which investigated the fusion of multiple information streams in the context of typical speech recognition. Building on this, in Yue et al. (2022d) multimodal acoustic-articulatory models for ADSR have been constructed and the impact of fusion at the low (input), medium (after convolutional layers) and high (right before output layer) levels were studied.

Fig. 9 illustrates our single and multimodal acoustic models. The architecture consists of a cascade of the convolutional, recurrent (LiGRU) and fully-connected (FC) layers. The nodes in the output layer represent state-clustered context-dependent (CD) triphones. In our multimodal systems, namely concat −1 and concat −2, modalities are fused at the input and medium levels, respectively. Notably, we did not explore information fusion at high levels. Previous work (Loweimi et al., 2020b, 2021a,b; Yue et al., 2022d; Loweimi et al., 2023b) consistently indicated that such fusion at high levels results in inferior performance.

## 3. Experimental results and discussion

### *3.1. Experimental setup*

The 83-D FBank features (80 log-FBank + 3 pitch-related features Ghahremani et al., 2014) are used as acoustic features in the baseline. EMA features are used as articulatory features. The training data is augmented using speed perturbation by the following factors: 0.9 (slower), 1.0 (original) and 1.1 (faster), resulting in a three-fold expansion of the training data. The used CNNs are

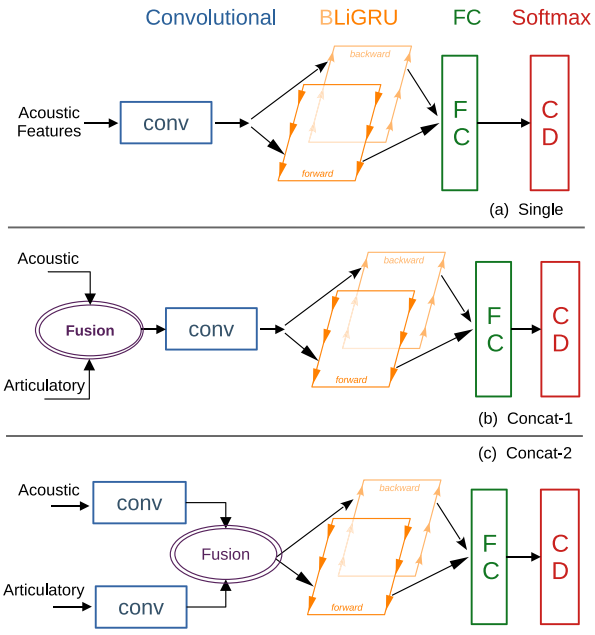



**Fig. 9.** The proposed multimodal acoustic modelling system, consisting of a cascade of convolutional, bidirectional light-gated GRU (BLiGRU) and fully-connected (FC) layers. The output layer is a state-clustered context-dependent (CD) triphones. (a) Single-modal baseline system, (b) multimodal system with fusion at the input level (concat −1), (c) multimodal system with fusion at the medium level (concat −2).

cascades of three 1-D convolutional layers, and they are followed by a stack of five bidirectional LiGRU (Ravanelli et al., 2018; Graves et al., 2013) layers with 550 units per direction, a fully-connected layer and one softmax classifier (estimating the context-dependent states). The dropout (0.15) (Srivastava et al., 2014), layer normalisation (Ba et al., 2016) and batch normalisation (Ioffe and Szegedy, 2015) are also applied along with RMSProp optimisation (Tieleman and Hinton, 2012). Learning-rate annealing is deployed with a factor of 0.5. The 5-fold cross-training setup proposed in Yue et al. (2020a) is applied for all experiments in this work. DNNs were trained by the PyTorch-Kaldi toolkit (Ravanelli et al., 2019; Paszke et al., 2017; Povey et al., 2011). An independent 200k vocabulary size Librispeech (Panayotov et al., 2015) trigram language model proposed in Yue et al. (2020) is employed for decoding. The acoustic models were trained with the state-clustered context-dependent triphones using the cross-entropy (CE) loss function.

For raw signal-based acoustic modelling, we employed raw waveform, raw magnitude spectrum (Loweimi et al., 2020b), raw magnitude spectra of the source (excitation) and filter (vocal tract) components (Loweimi et al., 2021c; Yue et al., 2022a,c) as well as the raw real and imaginary parts of the Fourier transforms (Loweimi et al., 2023b). For raw waveform modelling, we have used both non-parametric CNNs and SincNet (Ravanelli and Bengio, 2019) which is a parametric CNN. For further detail on acoustic modelling for each raw signal representation, readers are referred to the respective references.

### 3.2. Performance of the three articulatory measures

Table 5 compares the WER of systems trained on various input features and different articulatory measures. The training data is a combination of dysarthric and typical speech. The results are averaged across speakers in the dysarthric and typical groups separately. It is observed that both the FBank83+Tongue and FBank83+Lip systems outperform the baseline *FBank83* system, reducing WER by 0.2% and 0.6% (absolute).[4] on average for the dysarthric speech. The performance gains of 0.5% and 0.2% are also achieved for the typical speech by integrating the lip and tongue information. The lip information is more beneficial to the ADSR system than the tongue information. This is in agreement with the conclusion of the MI analysis in Section 2.4 which suggests that lip articulators would be more beneficial for improving the accuracy of ADSR. As a result, we decided to use the lip information in the following experiments.

The bottom block of Table 5 compares the results of combining FBank83 features with three different lip articulatory measures (*Lip, Lip_EuD_origin, Lip_EuD*), where EuD refers to Euclidean distance-based articulatory features. The results indicate that FBank83+Lip_EuD outperforms other articulatory measures. The application of EuD features reduces WER by 1.0% for both dysarthric and typical speech, compared to the baseline FBank83 system, yielding the highest performance gain relative to other

[4] $p<.05$. We performed statistical significance tests using the Matched Pairs Sentence-Segment Word Error (MAPSSWE) method (Gillick and Cox, 1989), following (WER Statistical Significance Test. [online]. Available: https://github.com/talhanai/wer-sigtest).

**Table 5**
WER for systems trained on various input features and different articulatory measures.

| Input Features | Average | |
|---|---|---|
| | Dysarthric | Typical |
| FBank83 | 35.6 | 11.7 |
| FBank83+EMA (12) | 35.4 | 11.7 |
| FBank83+Tongue | 35.4 | 11.5 |
| FBank83+Lip | **35.0** | **11.2** |
| FBank83+Lip+Tongue | 35.3 | 11.4 |
| FBank83+Lip | **35.0** | **11.2** |
| FBank83+Lip_EuD_origin | 34.7 | 11.0 |
| FBank83+Lip_EuD | **34.6** | **10.7** |

**Table 6**
WER for the baseline and different concatenation levels.

| Input Features | Fusion | Average | |
|---|---|---|---|
| | | Dys | Typ |
| FBank83 | **baseline** | 35.6 ± 3.3 | 11.7 ± 1.0 |
| FBank83+EMA | **concat-1** | 34.4 ± 2.2 | 10.8 ± 0.4 |
| FBank83+EMA | **concat-2** | **33.9 ± 2.0** | **10.4 ± 0.6** |

**Table 7**
WER for different features per (F)emale and (M)ale speakers with different dysarthria severity, along with the averaged results for all speakers. 'M/S' indicates speakers with Moderate to Severe levels of dysarthria.

| Input features | Severe | | | M/S | Moderate | Mild | | Average | |
|---|---|---|---|---|---|---|---|---|---|
| | M01 | M02 | M04 | M05 | F03 | F04 | M03 | Dys | Typ |
| FBank83 | 60.4 | 55.2 | 68.7 | 46.6 | 31.5 | 16.8 | 8.4 | 35.6 ± 3.3 | 11.7 ± 1.0 |
| Mag | 59.7 | 54.8 | 68.9 | 46.2 | 31.2 | 16.5 | 8.3 | 35.3 ± 2.7 | 11.5 ± 0.8 |
| VT+Exc | 58.3 | 54.8 | 66.5 | 46.3 | 30.8 | 16.4 | 8.3 | 34.3 ± 1.9 | 11.4 ± 0.3 |
| Real+Imag | 58.5 | 54.6 | 67.3 | 46.1 | 31.0 | 16.4 | 8.2 | 34.6 ± 2.1 | 11.3 ± 0.6 |
| Raw-CNN | 66.3 | 57.5 | 70.8 | 50.1 | 33.7 | 19.1 | 11.0 | 40.3 ± 3.8 | 13.7 ± 1.8 |
| SincNet | 58.5 | 56.0 | 67.2 | 46.3 | 31.0 | 16.1 | 8.3 | 35.5 ± 3.8 | 11.3 ± 1.8 |
| FBank83+EMA | 56.6 | 52.9 | 67.9 | 45.4 | 31.1 | 14.5 | 7.6 | 33.9 ± 2.0 | 10.4 ± 0.6 |
| Mag+EMA | 57.1 | 52.5 | 68.1 | 43.1 | 32.0 | 14.6 | 7.3 | 34.1 ± 2.1 | 10.4 ± 0.6 |
| VT+Exc+EMA | 55.9 | 52.1 | 66.1 | 42.0 | 30.3 | 14.2 | 7.3 | **32.9 ± 1.6** | **10.2 ± 0.4** |
| Real+Imag+EMA | 56.1 | 52.5 | 67.6 | 42.8 | 30.3 | 14.3 | 7.3 | 33.3 ± 2.0 | **10.2 ± 0.5** |
| Raw-CNN+EMA | 65.0 | 56.5 | 68.3 | 48.9 | 33.1 | 18.3 | 10.6 | 39.6 ± 3.3 | 13.1 ± 1.1 |
| SincNet+EMA | 56.8 | 52.3 | 66.3 | 44.8 | 30.5 | 14.6 | 8.0 | 33.1 ± 3.3 | 10.4 ± 1.1 |
| **Conformer** | | | | | | | | | |
| FBank83 | 68.1 | 60.7 | 71.3 | 62.3 | 17.6 | 11.3 | 5.4 | 37.1 ± 2.9 | 17.7 ± 0.9 |
| FBank83+EMA | 69.3 | 61.5 | 71.6 | 75.7 | 22.9 | 13.7 | 7.8 | 41.5 ± 2.7 | 17.1 ± 1.1 |

articulatory features. Accordingly, we will employ it (i.e., Lip_EuD) as an articulatory feature in the following experiments, and name it as "EMA" for convenience. We hypothesise that the Lip_EuD mitigates the influence of head movement and implicitly normalises the articulatory features, contributing to improved ADSR performance. Furthermore, mutual information analysis in Section 1.5 revealed that lip features are less distorted when comparing dysarthric and typical speech and exhibit higher mutual information with phone alignment labels.

### *3.3. Performance of different fusion schemes*

The results of testing the different feature fusion systems are reported in Table 6. Comparing the input and medium level concatenation: *concat −1* and *concat −2* with the baseline *FBank83* system shows 1.2% and 1.7% absolute WER reductions for dysarthric speech, respectively. The *concat −2* appears to be the best fusion scheme.

### *3.4. Raw signal acoustic modelling*

To explore the effectiveness of various acoustic features when fused with articulatory information, Table 7 reports WERs for systems using raw signal representations, namely raw magnitude spectrum, vocal tract (VT) and excitation (Exc), real (Real) and imaginary (Imag) components and raw waveform models with parametric and non-parametric CNNs (SincNet and Raw CNN). The consistent performance improvement observed after incorporating articulatory features suggests that integrating articulatory information consistently enhances the performance across all types of input features compared to solely relying on acoustic features.

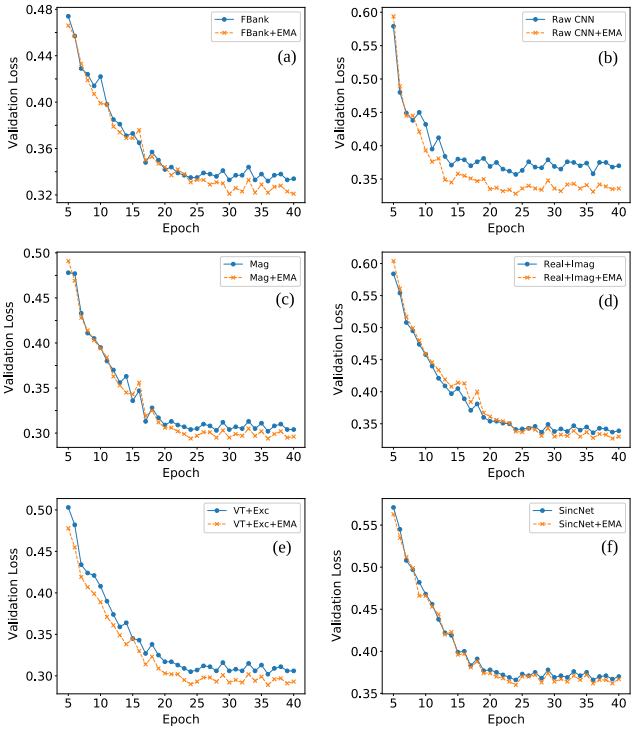



**Fig. 10.** Cross-entropy loss vs. epoch for systems with and without EMA. (a) FBank, (b) raw waveform with non-parametric CNN, (c) raw magnitude spectrum, (d) raw real (Real) and imaginary parts (Imag), (e) raw vocal tract (VT) and excitation (Exc) components, (f) raw waveform with SincNet.

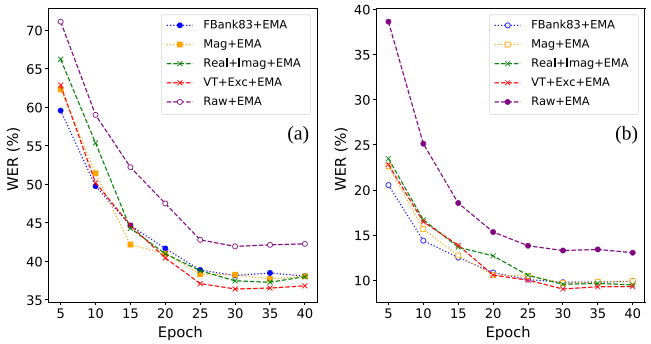



**Fig. 11.** WERs across different epochs for various systems evaluated on (a) speakers with dysarthria and (b) typical speakers. EMA: Lip_EuD.

The best performance is achieved by VT+Exc+EMA with 1.4% and 1.2% absolute WER reductions for dysarthric and typical speech compared with the VT+Exc system. The improvement indicates that some of the raw signal representations (e.g., VT+Exc, Real+Imag and Raw waveform with SincNet) are more effective when combined with articulatory information than hand-crafted features such as FBank.

For fair comparison, we also applied SOTA E2E Conformer[5] ASR model trained on the paired acoustic-articulatory data. As shown in the last two rows in Table 7, the performance was suboptimal compared to our hybrid model. This might be due to the limited amount of paired acoustic and articulatory data to train a sufficient Conformer model with the in-domain data only without using any out-of-domain data.

[5] The Conformer (Gulati et al., 2020) model used in this paper consists of 12 Conformer encoder layers and 6 Transformer decoder layers, both with output dimensions of 256. The attention mechanism uses 4 attention heads. The feed-forward layers have 1024 units in the encoder and 2048 in the decoder. The Conformer model was trained using a joint connectionist temporal classification (CTC)-attention objective (Kim et al., 2017), with a CTC weight of 0.3, an attention weight of 0.7, and 500 BPE units.

**Table 8**
The number of trainable parameters (in millions) for the models with different acoustic features, with (w.) and without (w.o.) EMA.

| | FBank83 | Mag | VT+Exc | Real+Imag | SincNet CNN |
|---|---|---|---|---|---|
| w.o. EMA | 10.1 | 9.8 | 10.0 | 10.0 | 10.1 |
| w. EMA | 10.3 | 9.9 | 10.2 | 10.2 | 10.3 |

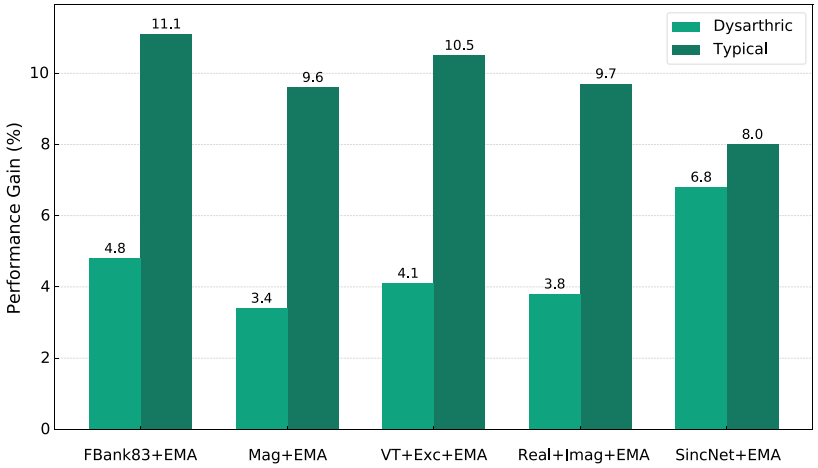



**Fig. 12.** Performance gain by feature type and speech type.

We also explored the training dynamics of the models. Fig. 10 depicts the training dynamics of various models in terms of the cross-entropy loss on the validation set in two conditions: with and without EMA information. The results indicate that incorporating EMA information does not affect the convergence rate but can slightly and consistently lower the cross-entropy loss by the end of training. We also explored the performance evolution of the systems of various features fusing with EMA lip information in terms of WER across different epochs. The results for speakers with dysarthric and typical speech are plotted in Fig. 11. As seen, the performance reaches a plateau after 25 epochs for dysarthric speech whilst for typical speech the performance keeps improving up to 40 epochs, which is consistent among different systems. VT+Exc+Lip starts to outperform other systems on dysarthric speech after 15 epochs' training. We counted the number of trainable parameters of models with different input features with and without EMA in Table 8. It is observed that fusing EMA information does not increase model complexity much but provides considerable improvement in recognition performance.

We also observe that the performance gain achieved by including the articulatory features varies depending on the speech types and the type of features. Fig. 12 represents the performance gain for different speech types. For dysarthric speech, the performance improvement varies, ranging from a minimum of 3.4% for the "Mag+EMA" features to a maximum of 6.8% for the "SincNet+EMA" features. In contrast, the performance improvement for typical speech is consistently higher across all features, reaching its peak at 11.1% for "FBank83+EMA" and dropping to 4.4% for "Raw-CNN+EMA". Fig. 12 also indicates that combining FBank and EMA data significantly enhances the recognition performance of both dysarthric and typical speech, followed by VT+Exc and Real+Imag. However, the Raw-CNN system shows the least improvement, particularly for dysarthric speakers. Larger performance gains are observed in typical speech compared to dysarthric speech across all feature types. This indicates that applying EMA data is more effective for typical speech. A possible explanation is that dysarthric speakers often have reduced control over their articulators, making the articulatory movements noisier and less consistent.

## 4. Conclusion

This paper demonstrated the effectiveness of incorporating real articulatory information along with raw acoustic features towards constructing multimodal acoustic-articulatory speech recognition systems. We illustrated the articulatory challenges that speakers with dysarthria have, as well as the motion mismatch between the dysarthric and typical speech in the articulatory space by visualising and systematically analysing the real articulatory data samples. This exploration provides evidence that due to articulation differences, compared with estimated articulation parameters from acoustic signals using knowledge about typical speech, the real articulatory data is more reliable, that is, able to represent the actual dysarthric speech articulatory space with less uncertainty. However, currently there is very little real articulatory data. Therefore, more data should be collected and researchers need to continue exploring optimal ways of using the available real/estimated articulatory data for ADSR.

We constructed effective multimodal dysarthric speech recognition systems by combining the acoustic and articulatory features with powerful acoustic models. The pair-wise Euclidean distance of the articulators in the lip region has been demonstrated to be the

most effective articulatory feature for the ADSR task. We also performed mutual information analysis, which revealed that lip-based articulatory features are more stable under dysarthric conditions and exhibit higher mutual information with phonetic alignment. These findings align with the ASR results, providing theoretical support for their effectiveness. Based on the observations, we made a suggestion that, for pathological speech data collection, video data capturing lip movements is a more effective and cost-efficient alternative to Electromagnetic Midsagittal Articulography (EMA) data.

Our proposed multi-stream acoustic models consist of convolutional, recurrent and fully-connected layers, allowing the multi-modal features to be fused via various schemes and at different levels of abstraction. The best performance was achieved by fusing the acoustic and articulatory information at the medium level trained on both dysarthric and typical data (FBank83+EMA), which reduces the absolute WER by 1.7% (4.8% relative reduction) on dysarthric speech compared with the FBank83 baseline system. Further improvement was achieved by exploiting the raw signal acoustic modelling, resulting in a 2.7% absolute (7.6% relative) WER reduction by using raw source and filter components (VT+Exc+EMA) compared with the FBank83 baseline. Notably, the performance improvement was more pronounced for severely dysarthric speech compared to mildly dysarthric speech.

Future work includes applying more advanced architectures and sophisticated acoustic-articulatory fusion schemes. Additionally, exploring the use of end-to-end models for the ADSR task, and transfer learning strategies with out-of-domain typical speech data, form other broad avenues for future research.

## CRediT authorship contribution statement

**Zhengjun Yue:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation. **Erfan Loweimi:** Writing – review & editing, Validation, Methodology. **Zoran Cvetkovic:** Supervision. **Jon Barker:** Supervision. **Heidi Christensen:** Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

I have shared the reference of the data in the paper.

## References

Ager, M., Cvetković, Z., Sollich, P., 2011. Combined waveform-cepstral representation for robust speech recognition. In: 2011 IEEE International Symposium on Information Theory Proceedings. IEEE, pp. 864–868.

Almadhor, A., Irfan, R., Gao, J., Saleem, N., Rauf, H.T., Kadry, S., 2023. E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. Expert Syst. Appl..

Ba, J., Kiros, J., Hinton, G., 2016. Layer normalization. In: Deep Learning Symposium. NIPS.

Badino, L., Canevari, C., F., L., Metta, G., 2016. Integrating articulatory data in deep neural network-based acoustic modeling. Comput. Speech Lang. 36, 173–195.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12449–12460.

Christensen, H., Aniol, M., Bell, P., Green, P., Hain, T., King, S., Swietojanski, P., 2013. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In: INTERSPEECH. pp. 3642–3645.

Darley, F., Aronson, A., Brown, J., 1969. Clusters of deviant speech dimensions in the dysarthrias. J. Speech Hear. Res. 12 (3), 462–496.

Duan, S., Zhang, X., Yan, M., Zhang, J., 2020. Statistical distribution exploration of tongue movement for pathological articulation on word/sentence level. IEEE Access 8, 91057–91069.

Duffy, J., 2013. Motor Speech Disorders-E-Book: substrates, Differential Diagnosis, and Management. Elsevier Health Sciences.

Espana-Bonet, C., Fonollosa, J., 2016. Automatic speech recognition with deep neural networks for impaired speech. In: International Conference on Advances in Speech and Language Technologies for Iberian Languages. Springer, pp. 97–107.

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., Benoit, A., 1995. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. Augment. Altern. Commun. 11 (3), 165–175.

Frankel, J., King, S., 2001. ASR-articulatory speech recognition. In: Seventh European Conference on Speech Communication and Technology.

Fujimura, O., 1986. Relative invariance of articulatory movements, in invariance and variability in speech processes. Lawrence Erlbaum 226–242.

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., Meng, H., 2022. Investigation of data augmentation techniques for disordered speech recognition. arXiv preprint arXiv:2201.05562.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., 2014. A pitch extraction algorithm tuned for automatic speech recognition. In: ICASSP. pp. 2494–2498.

Gillick, L., Cox, S.J., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: ICASSP. pp. 532–535.

Gowers, W., 2001. Clinical speech syndromes of the motor systems. In: Neurology for the Speech-Language Pathologist, fifth ed. Butter worth Heinemenn, Philadelphia, pp. 196–203.

Graves, A., Jaitly, N., Mohamed, A., 2013. Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, pp. 273–278.

Gulati, A., Qin, J., Wang, S., Zhang, Z., Wu, Y., Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. In: INTERSPEECH.

Hermann, E., Magimai.-Doss, M., 2020. Dysarthric speech recognition with lattice-free MMI. In: ICASSP. pp. 6109–6113.

Hill, D., Taube-Schock, C., Manzara, L., 2017. Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool. Can. J. Linguist. 62 (3), 371–410.

Hu, S., Xie, X., Geng, M., Jin, Z., Deng, J., Li, G., Wang, Y., Cui, M., Wang, T., Meng, H., et al., 2024. Self-supervised ASR models and features for dysarthric and elderly speech recognition. IEEE/ ACM Trans. Audio Speech Lang. Process..
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp. 448–456.
Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous univariate distributions, volume 2, vol. 289, John wiley & sons.
Joy, N., Umesh, S., 2018. Improving acoustic models in torgo dysarthric speech database. IEEE Trans. Neural Syst. Rehabil. Eng. 26 (3), 637–645.
Kent, R., 2000. Research on speech motor control and its disorders: A review and prospective. J. Commun. Disord. 33 (5), 391–428.
Kim, M., Cao, B., An, K., Wang, J., 2018. Dysarthric speech recognition using convolutional LSTM neural network. In: INTERSPEECH. pp. 2948–2952.
Kim, S., Hori, T., Watanabe, S., 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: ICASSP.
Kirchhoff, K., Fink, G., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. 37 (3–4), 303–319.
Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69, 066138.
Latha, M., Shivakumar, M., Manjula, G., Hemakumar, M., Kumar, M.K., 2023. Deep learning-based acoustic feature representations for dysarthric speech recognition. SN Comput. Sci. 4 (3), 272.
Loweimi, E., Bell, P., Renals, S., 2019. On learning interpretable CNNs with parametric modulated kernel-based filters. In: INTERSPEECH.
Loweimi, E., Bell, P., Renals, S., 2020a. On the robustness and training dynamics of raw waveform models. In: INTERSPEECH. pp. 1001–1005.
Loweimi, E., Bell, P., Renals, S., 2020b. Raw sign and magnitude spectra for multi-head acoustic modelling. In: INTERSPEECH. pp. 1644–1648.
Loweimi, E., Cvetkovic, Z., Bell, P., Renals, S., 2021a. Speech acoustic modelling from raw phase spectrum. In: ICASSP. pp. 6738–6742.
Loweimi, E., Cvetkovic, Z., Bell, P., Renals, S., 2021b. Speech acoustic modelling using raw source and filter components. In: INTERSPEECH. pp. 276–280.
Loweimi, E., Cvetkovic, Z., Bell, P., Renals, S., 2021c. Speech Acoustic Modelling Using Raw Source and Filter Components. In: INTERSPEECH. pp. 276–280.
Loweimi, E., Yue, Z., Bell, P., Renals, S., Cvetkovic, Z., 2023a. Multi-stream acoustic modelling using raw real and imaginary parts of the Fourier transform. IEEE/ ACM Trans. Audio Speech Lang. Process. 31, 876–890.
Loweimi, E., Yue, Z., Bell, P., Renals, S., Cvetkovic, Z., 2023b. Multi-stream acoustic modelling using raw real and imaginary parts of the Fourier transform. IEEE/ ACM Trans. Audio Speech Lang. Process. 31, 876–890.
Mitra, V., 2010. Articulatory information for robust speech recognition.
Mitra, V., Sivaraman, G., Bartels, C., Nam, H., Wang, W., Espy-Wilson, C., Vergyri, D., Franco, H., 2017. Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks. In: ICASSP.
Nam, H., Goldstein, L., Saltzman, E., Byrd, D., 2004. TADA: An enhanced, portable task dynamics model in MATLAB. J. Acoust. Soc. Am. 115 (5), 2430–2430.
Ouni, S., Mangeonjean, L., Steiner, I., 2012. VisArtico: a visualization tool for articulatory data. In: Thirteenth Annual Conference of the International Speech Communication Association.
Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books. In: ICASSP.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: NIPS Workshop on Autodiff.
Patel, H., Patil, H.A., 2024. Noise robust whisper features for dysarthric automatic speech recognition. Small 12 (768), 12.
Pedregosa, F., et al., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Povey, D., et al., 2011. The Kaldi speech recognition toolkit. In: ASRU.
Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, pp. 28492–28518.
Ravanelli, M., Bengio, Y., 2019. Speaker and speech recognition from raw waveform with SincNet. In: ICASSP.
Ravanelli, M., Brakel, P., Omologo, M., Bengio, Y., 2018. Light gated recurrent units for speech recognition. IEEE Trans. Emerg. Top. Comput. Intell. 2 (2), 92–102.
Ravanelli, M., Parcollet, T., Bengio, Y., 2019. The pytorch-kaldi speech recognition toolkit. In: ICASSP. pp. 6465–6469.
Ross, B.C., 2014. Mutual information between discrete and continuous data sets. PLoS ONE 9 (2), e87357.
Rudzicz, F., 2009. Applying discretized articulatory knowledge to dysarthric speech. In: ICASSP. IEEE, pp. 4501–4504.
Rudzicz, F., 2010a. Articulatory knowledge in the recognition of dysarthric speech. IEEE Trans. Audio Speech Lang. Process. 19 (4), 947–960.
Rudzicz, F., 2010b. Learning mixed acoustic/articulatory models for disabled speech. In: Proceedings of the Workshop on Machine Learning for Assistive Technologies. the Twenty-Fourth Annual Conference on Neural Information Processing Systems. NIPS, Citeseer, pp. 70–78.
Rudzicz, F., Hirst, G., van Lieshout, P., 2012a. Vocal tract representation in the recognition of cerebral palsied speech. J. Speech Lang. Hear. Res..
Rudzicz, F., Namasivayam, A., Wolff, T., 2012b. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. 46 (4), 523–541.
Sainath, T.N., Vinyals, O., Senior, A., Sak, H., 2015. Convolutional, long short-term memory, fully connected deep neural networks. In: ICASSP.
Schönle, P., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., Conrad, B., 1987. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang. 31 (1), 26–35.
Shahamiri, S.R., 2021. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Trans. Neural Syst. Rehabil. Eng. 29.
Shahamiri, S.R., Lal, V., Shah, D., 2023. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. IEEE Trans. Neural Syst. Rehabil. Eng..
Shahrebabaki, A., Olfati, N., Imran, A., Johnsen, M., Siniscalchi, S., Svendsen, T., 2021. A two-stage deep modeling approach to articulatory inversion. In: ICASSP. pp. 6453–6457.
Shahrebabaki, A., Siniscalchi, S., Salvi, G., Svendsen, T., 2020. Sequence-to-sequence articulatory inversion through time convolution of sub-band frequency signals. Database 1, 5.
Soleymanpour, M., Johnson, M.T., Soleymanpour, R., Berry, J., 2022. Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition. In: ICASSP. IEEE, pp. 7382–7386.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 1929–1958.
Takashima, Y., Nakashika, T., Takiguchi, T., Ariki, Y., 2015. Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition. In: EUSIPCO 2015. IEEE, pp. 1411–1415.
Takashima, R., Sawa, Y., Aihara, R., Takiguchi, T., Imai, Y., 2024. Dysarthric speech recognition using pseudo-labeling, self-supervised feature learning, and a joint multi-task learning approach. IEEE Access.
Teplansky, K., Wisler, A., Cao, B., Liang, W., Whited, C., Mau, T., Wang, J., 2020. Tongue and lip motion patterns in alaryngeal speech. In: INTERSPEECH. pp. 4576–4580.
Tieleman, T., Hinton, G., 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. In: COURSERA Neural Networks Mach. Learn.
Tüske, Z., Schlüter, R., Ney, H., 2018. Acoustic modeling of speech waveform based on multi-resolution, neural network signal processing. In: ICASSP.

Udupa, S., Siddarth, C., Ghosh, P.K., 2023. Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models. In: ICASSP. IEEE, pp. 1–5.

Vinotha, R., Hepsiba, D., Anand, L.V., 2024. Leveraging openai whisper model to improve speech recognition for dysarthric individuals. In: 2024 Asia Pacific Conference on Innovation in Technology. APCIT.

Wang, T., Hu, S., Deng, J., Jin, Z., Geng, M., Wang, Y., Meng, H., Liu, X., 2023a. Hyper-parameter adaptation of conformer ASR systems for elderly and dysarthric speech recognition. arXiv preprint arXiv:2306.15265.

Wang, H., Thebaud, T., Villalba, J., Sydnor, M., Lammers, B., Dehak, N., Moro-Velazquez, L., 2023b. Duta-vc: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model. arXiv preprint arXiv:2306.10588.

Wilson, B., Blaney, J., 2000. Acoustic variability in dysarthria and computer speech recognition. Clin. Linguist. Phon. 14 (4), 307–327.

Wrench, A., Richmond, K., 2000. Continuous speech recognition using articulatory data. In: Sixth International Conference on Spoken Language Processing.

Xiong, F., Barker, J., Christensen, H., 2018. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In: Speech Communication; 13th ITG-Symposium.

Yılmaz, E., Mitra, V., Bartels, C., Franco, H., 2018. Articulatory features for ASR of pathological speech. arXiv preprint arXiv:1807.10948.

Yılmaz, E., Mitra, V., Sivaraman, G., Franco, H., 2019. Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. Comput. Speech Lang. 58, 319–334.

Young, V., Mihailidis, A., 2010. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. Assist. Technol. 22 (2), 99–112.

Yousafzai, J., Sollich, P., Cvetkovic, Z., Yu, B., 2010. Combined features and kernel design for noise robust phoneme classification using support vector machines. IEEE Trans. Audio Speech Lang. Process. 19 (5), 1396–1407.

Yue, Z., Christensen, H., Barker, J., 2020a. Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition. In: INTERSPEECH.

Yue, Z., Loweimi, E., Christensen, H., Barker, J., Cvetkovic, Z., 2022a. Acoustic modelling from raw source and filter components for dysarthric speech recognition. IEEE/ ACM Trans. Audio Speech Lang. Process. 30, 2968–2980.

Yue, Z., Loweimi, E., Christensen, H., Barker, J., Cvetkovic, Z., 2022b. Dysarthric speech recognition from raw waveform with parametric CNNs. In: INTERSPEECH.

Yue, Z., Loweimi, E., Cvetkovic, Z., 2022c. Raw source and filter modelling for dysarthric speech recognition. In: ICASSP.

Yue, Z., Loweimi, E., Cvetkovic, Z., Christensen, H., Barker, J., 2022d. Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition. In: ICASSP.

Yue, Z., Xiong, F., Christensen, H., Barker, J., 2020. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In: ICASSP.