Human-Centric Intelligent Systems (2025) 5:103–122
https://doi.org/10.1007/s44230-025-00094-6

RESEARCH ARTICLE

Towards End-to-End Speech Articulation and Spoken Language Analysis Using Deep Learning

Tobias Weise[1,2] · Kubilay Can Demir[1] · Paula Andrea Pérez-Toro[2,4] · Tomas Arias-Vergara[2] · Andreas Maier[2] · Elmar Nöth[2] · Maria Schuster[3] · Björn Heismann[1] · Seung Hee Yang[1]

Received: 9 December 2024 / Accepted: 28 February 2025 / Published online: 14 March 2025

Abstract

This study presents a speech and spoken language analysis framework, leveraging a robust, end-to-end deep learning model developed in our prior work. The framework represents a foundational step towards a comprehensive solution for analyzing speech articulation and spoken language. Unlike traditional approaches that rely on separate specialized models, our architecture integrates multiple prediction tasks into a single multi-task learning setup: nine articulatory trajectories, a phoneme sequence, and phoneme alignment. While conceptually distinct, these outputs share a strong underlying relation: phonemes, as the fundamental building blocks of language, emerge from specific articulatory configurations, and phoneme alignment provides crucial temporal structure. We bridge the gap between abstract linguistic representations and their physical realizations by integrating phoneme recognition, articulatory trajectory prediction, and phoneme alignment within a single deep learning framework. Phonemes, as abstract speech units, manifest as concrete articulatory gestures, which can be precisely captured through EMA and analyzed using deep learning methods. This integration lays the foundation for diverse applications, including intelligibility assessment and therapeutic feedback. Extensive experiments validate the model's capabilities and demonstrate its potential in real-world contexts. These include evaluations of articulatory and phoneme-related metrics, intelligibility estimation using phoneme error rates, and open vocabulary keyword spotting. A case study on stroke-related datasets highlights how the framework provides detailed articulatory feedback and supports therapy progress tracking. While not a complete solution, this work shows that an integrated, end-to-end deep learning approach can effectively address multiple facets of speech analysis. Ultimately, it serves as a foundation for developing scalable and robust frameworks to tackle challenges in speech and language processing.

Keywords Acoustic to articulatory speech inversion · Dysarthria severity · EMA · Phoneme alignment · Tract variables · Wav2vec 2.0

✉ Tobias Weise
tobias.weise@fau.de

Kubilay Can Demir
kubilay.c.demir@fau.de

Paula Andrea Pérez-Toro
paula.andrea.perez@fau.de

Tomas Arias-Vergara
tomas.arias@fau.de

Andreas Maier
andreas.maier@fau.de

Elmar Nöth
elmar.noeth@fau.de

Maria Schuster
maschust@med.uni-muenchen.de

Björn Heismann
bjoern.heismann@fau.de

Seung Hee Yang
seung.hee.yang@fau.de

[1] Artificial Intelligence in Biomedical Engineering Department, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Bavaria, Germany

[2] Pattern Recognition Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Bavaria, Germany

[3] Department of Otorhinolaryngology, Head and Neck Surgery, Ludwig-Maximilians University, 80333 Munich, Bavaria, Germany

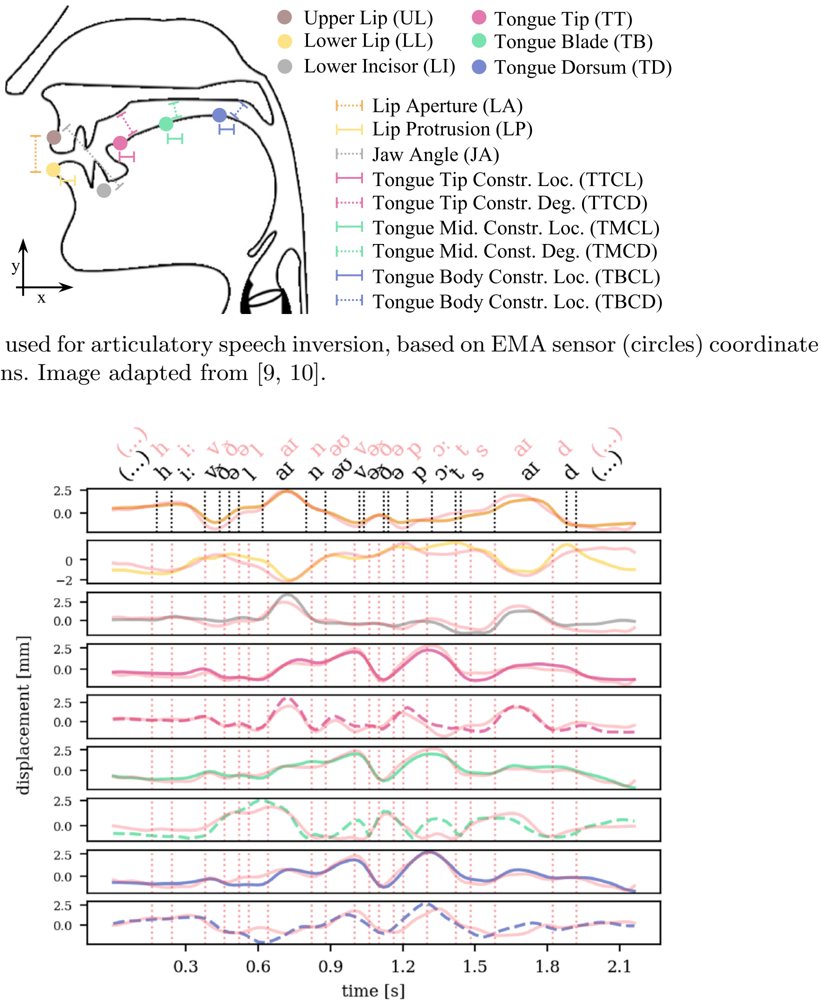
[4] GITA Lab, Faculty of Engineering, Universidad de Antioquia, Medellín 050010, Colombia

1 Introduction

Speech and spoken language are fundamental to human interaction, involving the production and perception of sounds to convey meaning. At the core of this process are speech units that serve as the building blocks of language. These units fall into two categories: abstract and concrete. This distinction is crucial for understanding how linguistic meaning is constructed and communicated through speech. Central to this system are the concepts of phonemes and phones, which are distinct units of human speech. Phonemes are the smallest meaningful sound distinctions in a specific language, functioning as abstract linguistic units that can alter the meaning of a word when substituted. In contrast, phones and their variant forms, allophones, are concrete speech units: physical realizations of phonemes that vary depending on speaker articulation and phonetic context. The relationship between phonemes and phones explains how speech is perceived and articulated, linking abstract linguistic structures to their articulatory and auditory expressions.

The parts of the vocal tract involved in creating sounds are called articulators, including the lips, jaw, tongue, and palate. Various techniques exist for measuring articulator movements [1], including X-ray, magnetic resonance imaging (MRI), and ultrasound. MRI provides excellent spatial resolution and is widely used for imaging vocal tract structures [2, 3]. However, its limited temporal resolution makes it less suitable for capturing rapid articulatory dynamics in real-time speech production. Ultrasound imaging is a non-invasive method that allows for real-time tongue tracking but lacks detailed articulatory information beyond lingual gestures [4, 5]. While ultrasound has been successfully applied in speech therapy and phonetics research, its use in large-scale automated speech analysis remains limited due to variability in image quality and tongue visibility constraints.

Fig. 1 The relation between EMA sensors, tract variables, and f-APTAI model predictions based on raw audio



(a) Nine TVs, used for articulatory speech inversion, based on EMA sensor (circles) coordinate transformations. Image adapted from [9, 10].

(b) Ground truth and `f-APTAI` predictions (light red color), of an unseen HPRC–N speaker (F03), uttering "heave the line over the port side": TV's, phoneme sequence, and alignment.

Electromagnetic articulography (EMA) (see Fig. 1a) offers a key advantage over these methods due to its high spatiotemporal resolution, enabling continuous, real-time tracking of articulatory movements with millimeter precision. Unlike MRI, which provides high spatial resolution but suffers from poor temporal resolution, or ultrasound, which is limited to tongue imaging, EMA allows for detailed, continuous monitoring of multiple articulators, making it particularly suitable for speech analysis tasks. These sensor coordinates are naturally speaker-specific, as they rely on the unique vocal tract anatomy of the individual being recorded.

So-called tract variables (TVs), introduced by Browman et al. [6], combine several vocal tract articulator movements into defined ”gestures” that achieve specific linguistic objectives relevant to articulation. These TVs provide a more abstract representation of articulatory movements, facilitating the modeling of articulatory-acoustic relationships. Figure 1a illustrates the relationship between EMA sensor coordinates and the resulting TVs. Ji [7] developed transformations to convert EMA sensor coordinates into TVs, making them less dependent on individual speaker characteristics compared to the original measurements [8]. Exemplary mathematical transformations from EMA signals to TVs are formally described in Eqs. (4) and (3), which define articulatory dimensions based on sensor positions.

Traditional approaches to speech and language analysis typically use specialized single-task models, ensemble methods, or signal-processing techniques. These methods assign distinct sub-tasks to individual components, achieving strong performance in specific areas but lacking cohesion. They often fail to fully exploit the interconnectedness of their components and face significant drawbacks, such as the need to manage distinct datasets for each task and the challenges of retraining or updating individual models to maintain compatibility. These limitations hinder scalability and adaptability, particularly in dynamic, real-world environments.

In this work, we propose an alternative: a unified framework for speech articulation and spoken language analysis, built around a robust end-to-end (E2E) deep learning model. This framework consolidates multiple sub-tasks into a single system, leveraging the inherent relationships between phonemes, articulatory trajectories, and phoneme alignment to provide a cohesive and holistic approach. By simplifying the analytical pipeline, this design enhances robustness and scalability, laying the foundation for applications such as second language learning and speech therapy.

This study represents a significant step towards a fully integrated framework. While not yet a complete solution, it demonstrates the potential of such an approach in addressing complex speech and spoken language challenges. We demonstrate the framework’s ability to analyze impaired intelligibility and provide actionable insights for therapeutic progress tracking, using stroke survivor therapy as an example. The framework builds on acoustic phoneme-to-articulatory inversion (APTAI), introduced in our prior work [9], which unifies the previously separate tasks of acoustic-to-articulatory inversion (AAI), phoneme-to-articulatory motion estimation (PTA), phoneme recognition, and phoneme alignment.

Our prior work also introduced the `f-APTAI` model, which serves as the cornerstone of this framework. This model outperformed a simpler variant in cross-corpus evaluations, validating its suitability for real-world applications. In a multi-task learning setup, it leverages a self-supervised foundation model, fine-tuned on diverse real-world datasets containing natural noise sources, including background environmental sounds, microphone distortions, and pronunciation variations. Rather than relying on preprocessing to remove these artifacts, the model is designed to learn from them, enhancing robustness for speaker- and text-independent predictions. To contextualize the framework, we focus on speech and language impairments in stroke survivors as a case study. The datasets include patients with stroke-related speech disorders, such as dysarthria, while the experiments focus on language disorders, such as aphasia. These evaluations demonstrate the framework’s ability to analyze impaired intelligibility and support therapeutic progress tracking.

1.1 Related Works

This section presents related research in areas relevant to this study. For related work on EMA, TV, and phoneme alignment, refer to our prior study [9], which details the `f-APTAI` deep learning model for the APTAI task. Cummins et al. [10] surveyed approaches in speech-based health detection, emphasizing the impact of deep learning. They concluded that although deep learning is increasingly used in speech-based health detection, it has yet to achieve the dominant influence seen in other fields. They also suggested future research directions to fully exploit its advantages. Yang et al. [11] reviewed research on deep learning for Alzheimer’s disease detection via speech analysis, highlighting its advantages for large-scale screening over traditional methods such as electroencephalography (EEG) and MRI.

Traditional approaches to speech and language analysis typically rely on single-task models, ensemble methods, or classical signal-processing techniques. These models are often optimized for specific sub-tasks such as phoneme recognition [12], articulatory trajectory prediction [13], or dysarthric speech classification [14]. While achieving high performance in isolated domains, they frequently lack cohesion and fail to fully exploit the inherent relationships between speech components, such as phoneme-articulatory mapping. This limitation has been recognized in prior research on articulatory speech modeling [15] and phoneme-to-speech conversion frameworks [16]. Additionally, updating or adapting these models to new datasets can be computationally expensive, as highlighted in [17].

In contrast, multi-task learning architectures have demonstrated significant advantages by simultaneously learning multiple speech-related tasks, leading to improved generalization and robustness [17, 18]. Recent studies have also highlighted how pre-trained self-supervised learning (SSL) models can improve feature extraction and enable adaptation across different speech domains [18]. Our proposed approach builds on this trend by integrating articulatory trajectory prediction, phoneme recognition, and phoneme alignment into a unified deep learning model.

1.2 Contributions

This work proposes a unified deep learning-based framework to address the complexity of speech articulation and spoken language analysis, marking a step toward scalable and integrated solutions. By combining phoneme recognition, articulatory trajectory prediction, and alignment within a single system, the framework bridges the gap between isolated tasks and holistic analysis. Beyond individual components, this study lays a foundation for advanced applications, including speech therapy and intelligibility assessment in health-related contexts. The design emphasizes adaptability and extensibility, showcasing its potential as a blueprint for similar approaches in related domains. The main contributions of this paper are as follows:

- We propose a novel framework for speech articulation and spoken language analysis, based on a unified, E2E deep learning model that integrates multiple outputs into a single system. This design leverages the strong relationships among outputs to lay an analytical foundation for diverse applications.
- Our model (`f-APTAI`), introduced in prior work [9], demonstrates competitive, state-of-the-art (SOTA) performance in articulatory trajectory prediction, phoneme recognition, and alignment, enabling robust analysis of speech and language tasks.
- Using stroke survivor therapy as a case study, we validate the framework through extensive experiments, including articulatory and phoneme metric evaluation, intelligibility estimation, and open vocabulary keyword spotting. These results demonstrate its ability to analyze impaired intelligibility and support therapeutic progress tracking.
- This work represents a significant step toward an integrated, E2E deep learning framework for multi-faceted speech and spoken language analysis, with broad potential for research and practical applications.

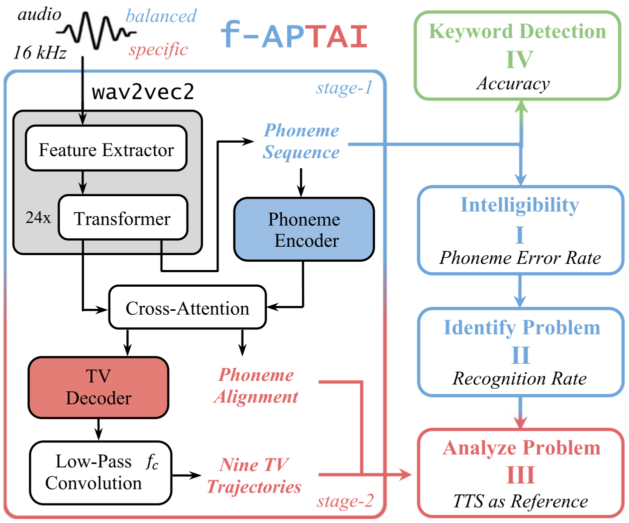


Fig. 2 Proposed speech and spoken language analysis framework, with the `f-APTAI` model's architecture on the left-hand side. The right-hand side shows different analysis steps I–IV, which will be referenced throughout the article

2 The f-APTAI Model

This section details the `f-APTAI` model [9], the core component of our speech and spoken language analysis framework. Our code is available online,[1] and a simplified architectural overview is shown on the left-hand side of Fig. 2. The following subsections describe the foundational architecture (Sect. 2.1). Next, we provide a brief overview of the two training stages (Sects. 2.2 and 2.3) of our model. The first stage fine-tunes the foundational model for robust and accurate phoneme recognition. Consequently, the learned weights are frozen during the second training stage, where an alignment matrix is trained alongside a TV regression task. For further technical details and methodological rationale, refer to our previous work [9]. Once trained, the model produces three outputs for a given input audio file: a frame-asynchronous phoneme sequence, a frame-synchronous phoneme alignment, and nine frame-synchronous TV predictions, all based on the `wav2vec2` feature encoder.

As an alternative, we experimented with specialized attention mechanisms for articulatory prediction, specifically the Feed-Forward Transformer (FFT) from Microsoft's FastSpeech [19]. We selected FFT over a standard Transformer for its non-autoregressive nature, enabling parallel output prediction and improving computational efficiency for speech-related tasks. However, our results indicated that while FFTs work well for sequence-to-sequence tasks like speech synthesis, they struggled to maintain smooth

[1] https://github.com/tobwei/APTAI

temporal continuity in articulatory trajectory prediction. In contrast, LSTM-based architectures demonstrated superior performance in capturing the sequential dependencies and gradual articulatory movements required for accurate prediction. Given these findings, we opted for an LSTM-based approach in our final framework, specifically in the TV decoder block.

2.1 Wav2Vec 2.0 Foundation Model

We selected Wav2Vec 2.0 as the central feature extractor for our `f-APTAI` model due to its state-of-the-art (SOTA) performance in automatic speech recognition (ASR), making it particularly effective for phoneme recognition. By leveraging its SSL capabilities to extract rich acoustic representations, it provides robust and adaptable feature extraction. Additionally, its open-source availability ensures accessibility, adaptability, and ease of integration into our framework.

This architecture [20] consists of three main parts: a convolutional feature encoder, a transformer context encoder, and a vector quantization module. The vector quantization module is used only during self-supervised pre-training (contrastive optimization with vast amounts of unlabeled data), where the resulting weights extract meaningful and distinct speech units. The pre-trained weights of the first two components can then be further optimized (fine-tuned) for specific downstream tasks. The following describes these two components, which are then also relevant for our `f-APTAI` model.

The local feature extractor first encodes the raw audio waveform input, normalized to zero mean and unit variance, into vector representations on a discrete time scale, with an output frequency of 49 Hz. This is implemented as seven blocks, each containing 1-dimensional temporal convolutions, followed by layer normalization and GELU activation [21]. The convolutional kernel strides and widths are parameterized to produce a frame hop of approximately 20 ms and a window size of 25 ms, mimicking traditional audio signal processing methods such as MFCCs. In the large `wav2vec2` variant, the feature dimension is linearly projected from 512 to 1024 with dropout applied but no activation, producing features of size $1024 \times T$. Here, $T = \left\lfloor \frac{L}{1/49} \right\rfloor$ represents the number of time frames, where L is the total duration of the input audio in seconds and the time frames are produced at a rate of 49 Hz.

Multiple transformer encoder blocks process the local vector representations, capturing relationships between speech units across the input sequence using self-attention. The first step is to compute a relative positional encoding. Here, `wav2vec2` uses the relative position of prior extracted local vector representations, instead of adding a pre-computed absolute positional encoding (e.g., sinusoidal positional encoding [22]). This is achieved via a convolutional layer, with a large kernel size of 128, a stride of 1, padding of 64, and 16 groups, followed by GELU activation. This extends the receptive field of each speech unit's local vector representation from 20 ms to $128 \times 20\,\text{ms} = 2.5\,\text{s}$ by adding the computed relative positional embedding. Finally, layer normalization and dropout are applied, before feeding the embeddings into the 24 transformer layers with 16 self-attention heads.

2.2 f-APTAI Model Training: Stage-1

In the first stage of the `f-APTAI` model, we train a phoneme recognizer to predict a frame-asynchronous phoneme sequence for a given input audio file by fine-tuning pre-trained weights of a self-supervised architecture. In our previous work, we evaluated different architectures and weight initializations, finding that `wav2vec2-large-robust` performed best. Here, `wav2vec2` refers to an architecture (see Sect. 2.1), and `large-robust` refers to weights, which were the result of pre-training this architecture with a specific dataset. These weights can be fine-tuned for multiple downstream tasks, with ASR as the original intended use case. Our fine-tuning objective is phoneme recognition, which, unlike ASR, identifies individual speech sounds (phonemes) rather than transcribing spoken language into text for applications like virtual assistants or transcription services.

We fine-tune the `wav2vec2-large-robust` feature encoder and transformer layers by adding a linear layer to the final (24th) transformer encoder output and optimizing a connectionist temporal classification (CTC) loss [23]. An input speech signal first passes through the `wav2vec2` feature extractor (frozen, pre-trained weights), then the transformer layers (fine-tuned, pre-trained weights) learn the context between frames, and finally, a randomly initialized linear layer with 46 hidden units is added. This layer represents the 45 chosen phonemes (see Table 1) and includes a blank token for each `wav2vec2`-based frame at 49 Hz. This token is part of the CTC optimization, originally designed for ASR. It addresses the challenge of having variable-length input and output sequences (audio vs. text) where the alignment between the two sequences is unknown, which is solved by this token and a method to calculate the probability of all possible alignments. This loss optimization behaves like a state machine, similar to hidden Markov models (HMMs), where it only requires the phonemic transcription (instead of text) as additional input (besides audio) during training. However, the benefit of not requiring an alignment as input, also means that CTC does not produce an alignment as output. Rather, it outputs a frame-asynchronous phoneme label sequence through a frame-synchronous decoding procedure, using the blank token and multiple

Table 1 International phonetic alphabet (IPA) based phonemes and resulting groups considered for this study

Manner of articulation	
Stop	/p/, /t/, /k/, /b/, /d/, /g/
Nasal	/n/, /m/, /ŋ/
Trill	/r/
Fricative	/s/, /ʃ/, /z/, /f/, /h/, /ð/, /ʒ/, /θ/
Approximants	/j/, /w/
Lateral	/l/
Vowel	/aɪ/, /aʊ/, /e/, /eə/, /eɪ/, /iː/, /uː/, /æ/, /ɑː/, /ɒ/, /ɔɪ/, /ɔː/, /ə/, /əʊ/, /ɜː/, /ɪ/, /ɪə/, /ʊ/, /ʊə/, /ʌ/
Place of articulation	
Labial	/p/, /b/, /m/, /f/, /v/, /w/
Alveolar	/t/, /d/, /n/, /tʃ/, /dʒ/, /ʒ/
Velar	/k/, /g/, /ŋ/, /w/
Palatal	/j/, /tʃ/, /dʒ/
Postalveolar	/ʃ/
Central	/eə/, /eɪ/, /ɑː/, /ə/, /əʊ/, /ʌ/
Front	/aɪ/, /e/, /iː/, /æ/, /ɜː/, /ɪ/, /ɪə/
Back	/aʊ/, /uː/, /ɒ/, /ɔɪ/, /ɔː/, /ʊ/, /ʊə/
Voicing	
Voiceless	/p/, /t/, /k/, /ʃ/, /s/, /h/, /tʃ/, /θ/
Voiced	/m/, /n/, /b/, /d/, /g/, /dʒ/, /w/, /ð/, /ʒ/

possible alignment paths. Decoding uses the beam search algorithm with a beam width of 10.

2.3 f-APTAI Model Training: Stage-2

In the second stage of `f-APTAI` training, the objective is to learn an alignment matrix between the embedded phoneme sequence and the corresponding transformer embedding, maintaining the same time resolution as the feature encoder's output. For a given audio input, two outputs from the trained and frozen stage-1 model are used in stage-2: the predicted CTC-based phoneme sequence (serving as the upper bound) and the corresponding final transformer encoder hidden representation. Furthermore, multi-task learning is employed since the model has to simultaneously learn this alignment matrix and a TV regression task.

At the core of stage-2 training lies a cross-attention layer, which uses the attention mechanism [22] to learn an alignment matrix. This is inspired by [24] and utilizes only the key (K) and query (Q), but no value (V) matrices in the process. If an alignment is expressed as a matrix, it must be monotonic and diagonal, which is enforced using the forward-sum (FS) loss from hidden Markov models (adopted from [25, 26]). This contributes to the MTL optimization, which we use during the stage-2 training of our model. In addition to the alignment matrix, the cross-attention layer generates a hidden representation used as input for the TV decoder in the `f-APTAI` architecture. This decoder mainly uses auto-regressive layers to output nine tract variables for each `wav2vec2` based time-step. Finally, the decoder output passes through a single 1D convolutional layer with fixed weights, acting as a low-pass filter adapted from [27]. This improves the smoothness of predicted TV trajectories and facilitates weight optimization during backpropagation, enhancing overall regression performance. Equation (1), where N is the size of the Hanning window and $f_c, \forall n \in [0, N-1]$ is adapted from [27]. It describes this filter, where the first term corresponds to a Hanning window and the second term is a shifted (by $2\pi f_c$) and scaled sinc function, corresponding to the impulse response of the ideal low-pass filter with a cutoff frequency of f_c.

$$w(n) \propto \left(1 - \cos\left(2\pi \frac{n}{N-1}\right)\right) \operatorname{sinc}\left(2\pi f_c\left(n - \frac{N-1}{2}\right)\right) \tag{1}$$

The final TV regression part that contributes to the MTL optimization goal is the reconstruction mean square error (MSE) loss between the (smooth) predicted and ground truth TV values.

3 Proposed Framework

We explore two ways to incorporate our `f-APTAI` model into our proposed speech and spoken language analysis framework. These procedures form the basis of our experiments to validate their effectiveness (see Sect. 5). First, its hidden features can be extracted for downstream tasks incorporating alignment and TV information, linking the abstract structure of speech to its concrete articulatory execution. Second, model predictions can be used as a pipeline (see the right-hand side of Fig. 2), structured into four steps (I–IV) based on the model's three outputs. This structured approach ensures scalable and interpretable speech articulation and spoken language analysis, providing a unified framework that overcomes the limitations of separate task-specific models. The remainder of this section explains these concepts, also referenced in the experiments. Figure 2 separates the model outputs by training stage for visual clarity; however, the `f-APTAI` model is always trained in two stages. Given a raw audio input, the model predicts three outputs: a phoneme sequence (stage-1), nine TV trajectories, and a phoneme alignment (both stage-2).

3.1 Step I

This step uses the phoneme error rate (PER) as a metric (see Sect. 4.2) to assess speech intelligibility. This metric reflects how understandable a person's speech is, independent of background noise. Previously, word error rate

(WER) from ASR systems has been used successfully [28], but PER offers a finer-grained alternative, providing a more precise metric. Computing PER for an utterance requires converting its graphemes to phonemes, using those defined in the `f-APTAI` model (see Table 1). We used a state-of-the-art text-dependent forced aligner for this conversion (see Sect. 5). In our experiments (see Sect. 5.4), we examined whether PER correlates with human expert intelligibility scores. A key consideration is using a phonetically balanced set of utterances to accurately represent user intelligibility. Additionally, we evaluate whether a practical (i.e., minimal) number of utterances is sufficient for this step. In a therapeutic setting, a user (e.g., a dysarthric patient) could iterate through different utterance sets, receiving a PER/intelligibility score after each exercise. These scores can be tracked over time to monitor progress.

3.2 Step II

While Step I may suffice for some use cases, a more detailed analysis can be more effective and instructive. Step II aims to identify specific issues that PER alone cannot detect, enabling detailed analysis and feedback in Step III. Therefore, we propose using the recognition rate (see Sect. 4.2) of the `f-APTAI` predicted phoneme sequence for articulation-based phoneme groups (see Table 1), such as nasals or fricatives. Rather than analyzing entire phonetic groups, or after identifying a problematic group, individual phoneme recognition rates can be examined to pinpoint specific phoneme difficulties. Taking dysarthria as an example, a patient typically does not struggle with specific phonetic groups but rather experiences difficulties across all phonemes due to the disorder's global impact on the control of articulators. This is true for more severe cases, while mild to medium patients can typically suffer from consonant connection problems. Here, the latter could be detected by Step II of the proposed framework, while for the severe cases, working with the PER as a surrogate for intelligibility during Step I would be the most effective approach.

3.3 Step III

Step III requires identifying a specific issue in Step II, such as difficulties with nasal phoneme production. Step III provides a detailed analysis of problem-specific utterances using a text-to-speech (TTS)-generated reference version. When using problem-specific utterances (e.g., nasal ones) as input for the `f-APTAI` model, this step will utilize the predicted nine TV trajectories, in combination with the phonetic alignment. This combination can provide valuable insights into specific articulation errors made by a user/patient, which is expected to be the underlying cause of the not-recognized phonemes and thus affected phoneme groups. Here, it is beneficial to have a reference for the correct articulation of an utterance and the contained phonemes, as well as co-articulation. We find that a TTS-generated version works better and requires less effort, compared to using a control speaker (see Experiment 5.6). This detailed analysis could also be used by an (automated and digital) therapist, while the user receives a more accessible explanation. Moreover, the TTS version can not only be used as a visual reference but also as acoustic feedback, by playing a user the TTS and his own (likely erroneous) version of a problem-specific utterance, which can aid in the understanding of the problem.

3.4 Step IV

In the context of stroke-related disorders, this step can support therapy for patients with aphasia. This part of the framework performs keyword detection using a previously recognized phoneme sequence and a grapheme-to-phoneme conversion of the target keyword. Unlike traditional keyword detection, which relies on pre-defined words and specially labeled training data, the proposed method detects arbitrary keywords (i.e., open vocabulary keyword spotting) by comparing a recognized phoneme sequence to the target keyword's phoneme representation. However, stroke patients often experience both aphasia and dysarthria. Here, the proposed method assumes the patient has improved dysarthric speech to a relatively typical intelligibility level. Otherwise, keyword detection based on phoneme recognition may be error-prone. In the proposed framework, intelligibility can be improved through iterative speech exercises and analysis of Steps I–III, depending on patient progress

3.5 Hidden Features

In addition to Steps I–IV (Fig. 2), we propose using two hidden representations from the frozen `f-APTAI` model (Fig. 3). These embeddings serve as features for downstream speech and language tasks, with experiments conducted in a dysarthric setting. The first embedding, the *"TV embedding"*, is extracted from the bidirectional long short-term memory (LSTM) layer. This embedding has a shape of $512 \times T$, capturing articulation-related information from the LSTM's forward and backward computations. The second hidden representation, the *"alignment embedding"*, is derived from the cross-attention layer at the core of stage-2 training in the `f-APTAI` model. It combines phonemic information ($128 \times T$) with positional data ($N \times 128$), forming an embedding of shape $256 \times T$, where N is the maximum predicted phoneme sequence length.

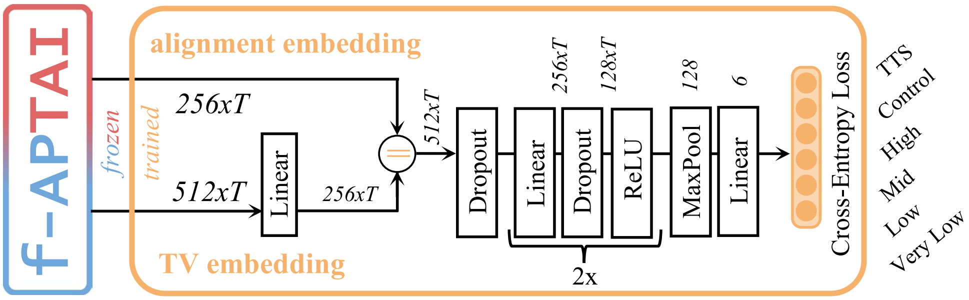


Fig. 3 Basic multi-class classification setup, showcasing how `f-APTAI`'s extracted hidden features can be used for the downstream task of dysarthria severity estimation

4 Data and Metrics

This study focuses on speech and language impairments following stroke, including aphasia and dysarthria. We limit our dataset selection to this context to maintain a well-defined experimental scope, ensuring consistency in pathological characteristics and model evaluation. This section describes the datasets and metrics we used in the experiments conducted for this study. This section details the datasets and metrics used in our experiments. We use four datasets and five metrics across seven experiments (see Sect. 5) to demonstrate use cases of the proposed speech and language analysis framework.

4.1 Data

Common Phone

For stage-1 training of `f-APTAI`, we use the Common Phone (CP) dataset [29] to fine-tune an English phoneme recognizer. It is derived from Mozilla's crowd-sourced Common Voice [30] corpus. We utilize the English subset containing 45 phoneme labels (see Table 1). Our primary motivation for using CP is to develop a robust system. This robustness is highlighted when comparing CP to datasets like TIMIT [31]. TIMIT recordings are made in a consistent, acoustically controlled environment with professional equipment, whereas CP recordings are collected from people's smartphones in various uncontrolled environments. A phoneme recognizer fine-tuned with such a dataset (recordings sampled at 16 kHz, matching `wav2vec2`'s requirement) learns from real-world conditions, including environmental background noise, varying microphone qualities, and spontaneous pronunciation inconsistencies. By directly modeling these variations rather than relying on artificial noise reduction, the approach enhances robustness and improves generalization across diverse speech settings.

HPRC

This is one of two datasets in this study that contain articulator-related information, specifically EMA sensor data recorded in parallel with acoustic data during spoken utterances. These sensors were placed on the tongue [tip (TT), blade (TB), rear (TR)), lips (upper (UL) and lower (LL)], left mouth corner (ML), and jaw (JAW). We use this dataset during stage-2 of the `f-APTAI` training and to evaluate speaker-independent performance. This dataset, the Haskins production rate comparison (HPRC) [32], contains recordings from four female and four male subjects reciting 720 phonetically balanced IEEE sentences at "normal" (HPRC-N) and "fast" (HPRC-F) speaking rates. Although the speakers repeat utterances, we randomly select only one repetition per utterance and speaker. We used a state-of-the-art text-dependent forced aligner to generate ground truth phoneme labels and time steps, ensuring compatibility with the CP dataset since the original dataset used a different forced aligner. To pre-process the EMA data, we addressed *NaN* values in some coordinates by applying linear interpolation, followed by low-pass (Butterworth) filtering at 20 Hz to eliminate recording-related noise. After this, the EMA coordinates were transformed into nine TVs (see Fig. 1a), and additional processing was applied: the original EMA data was sampled at 100 Hz, resulting in TVs at the same rate, which we resampled to 49 Hz to synchronize with the output frame rate of `wav2vec2`. Finally, we applied utterance-wise z-score normalization to the individual TVs.

TORGO

The second dataset, TORGO [33], contains EMA sensor data but lacks a palate trace. Previous research [34] proposed approximating the missing palate trace using the convex hull of tongue data, but we did not include this method in our experiments. Thus, our experiments included only six of the nine TVs, excluding tongue-related constriction degrees ("CD"). TORGO includes control and dysarthric speech recordings with aligned acoustic data and measured 3D articulatory features. Speakers complete tasks such as reading sentences, repeating words, and engaging in spontaneous speech. To the best of our knowledge, this is one of the few datasets containing EMA data for pathological speech. The sensor names differ from those in HPRC, with the following mappings: (HPRC, TORGO) → (JAW, LI), (TB, TM), (TR, TB). These mappings were applied in TV computations (see Sect. 5). The EMA data was preprocessed similarly to HPRC (see Sect. 4.1), except for its 200 Hz sampling rate, which was resampled to 49 Hz to match `wav2vec2`'s output frequency.

UASpeech

The UASpeech corpus [35] consists of audio recordings from 15 speakers with Cerebral Palsy (4 female, 11 male) and 13 healthy controls (4 female, 9 male). Additionally, it includes speaker-level expert intelligibility scores for dysarthric speakers, ranging from 2 to 95%. Recordings contain specific utterances (words) from each speaker, structured in common words (CW) and uncommon words (UW), with CW being repeated in each of the three recording blocks B1–B3. We select all CW from B1 and UW from all three blocks, resulting in a total of 455 unique utterances. After discarding corrupted utterances, we retain the same 448 utterances per speaker. Each session is recorded using an 8-channel microphone array. To obtain a single file per utterance and speaker, we randomly sample one of the microphone recordings on an utterance basis and resample it to 16 kHz. Furthermore, we use the same forced aligner to perform a grapheme-to-phoneme conversion of the utterance transcripts from this corpus. We also use the transcript provided by UASpeech to create TTS versions of the utterances for two male and two female artificial speakers (see Sect. 5).

4.2 Metrics

Phoneme error rate

The phoneme error rate (PER) is a ratio, computed by $PER = (D + I + S) / N$ where D number of deletions, I number of insertions, S the number of substitutions, and N is the total number of phonemes in the reference sequence. For this work, this reference is a grapheme-to-phoneme conversion result, based on the transcript of an utterance. Furthermore, the Levensthein distance algorithm computes the edit distance (minimum number of insertions, deletions, and substitutions) needed to convert the predicted sequence into the reference sequence. This error rate is then normalized by dividing the total number of edits, by the total number of phonemes in the reference sequence.

Pearson correlation coefficient

The Pearson correlation coefficient (PCC) is a unitless measure of the linear relationship between two quantities. It quantifies the degree to which two signals are related to each other. In related works regarding the AAI and PTA tasks, this metric is referred to as PCC, while it is traditionally noted as r. It is a coefficient that ranges from $[1, -1]$, where 1 indicates a perfect positive, -1 indicates a perfect negative, and 0 no linear relationship. It can be computed between two (e.g., TV) signals X and Y via

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2}$$

where X_i, Y_i are the i-th elements, and $\bar{X}, \bar{Y}$ are the mean of the respective signals, both having n elements. This indicates that it is computed element-wise, thus both signals must have the same length for the PCC to be computable.

Root mean square error

The root mean square error (RMSE) is a difference measure between the values, predicted by a model and the ground truth values. It is commonly used to evaluate the accuracy of a model, also in the context of related AAI and PTA tasks. Given two (e.g., TV) signals X and Y it can be computed by $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2}$, where X_i and Y_i are the i-th elements of the signals, and n is the number of elements in each signal, again indicating that it is computed on a per-element basis, requiring both signals to be of equal length.

Overlap

Overlap, measured as a percentage, quantifies the agreement between predicted and ground-truth phoneme sequences. Predictions from the `f-APTAI` model are compared against our upper bound for phoneme-related metrics. The upper bound is defined by a state-of-the-art text-dependent forced aligner with a Web-API (see Sect. 5). This ensures consistent phoneme labels across datasets, enabling direct comparisons (see Sect. 4).

Classification metrics

Some experiments are evaluated using classification metrics and additionally visualized in a confusion matrix. We use recognition rate (RR), also known as accuracy, along with recall and F1-score. The latter balances precision and recall, making it especially useful for imbalanced class distributions compared to accuracy.

5 Experiments

This section describes our seven conducted experiments, where we describe common facts in this paragraph, before going into detail about specific setups. Experiments two to seven are based on the first experiment, where we train and evaluate the `f-APTAI` model. Our models were implemented in PyTorch and ran on a single NVIDIA RTX 3090 TI 24GB GPU. Furthermore, the `wav2vec2` model and pre-trained weights were acquired from the Huggingface Transformers library. The state-of-the-art [24] text-dependent forced aligner that we use to create our ground-truth (thus upper bound) phoneme labels is called "WebMAUS" [36], which can be accessed via Web-API. Lastly, for TTS audio file generation we make use of OpenAI's text-to-speech API. The formulas that we use to convert the EMA sensor data to TVs are taken from [37], with two examples (LP, LA) shown below for better understanding:

$$LP[n] = LL_x[n] - \underset{m \in allutterances}{median} LL_x[m] \tag{3}$$

$$LA[n] = \sqrt{(LL_x[n] - UL_x[n])^2 + (LL_z[n] - UL_z[n])^2} \tag{4}$$

5.1 APTAI Performance

This is the experiment, intended to show the performance of our model in terms of its prediction outputs in the context of the APTAI task. This is important, as it is at the center of our proposed analysis framework. Training of both `f-APTAI` stages utilize the Adam optimizer, an initial learning rate of 1e−5, and a learning rate scheduler using warm-up, static, and decaying epochs.

For fine-tuning a phoneme recognizer during stage-1 of model training, we select wav2vec2-large-robust weights for their best overall performance, determined in our previous work. Furthermore, we utilize the CP dataset with its official train/dev/test splits, a batch size of 2, 160 epochs, a learning rate of 5e−6, a final dropout of 10%, and model selection based on validation PER.

For stage-2, we use the HPRC-N dataset (excluding HPRC-F to prevent PER degradation in stage-1) and evaluate performance using a leave-one-speaker-out (LOSO) approach. In this setup, data from seven speakers is allocated for training and validation (90%/10%), while data from the remaining speaker is reserved for testing, separated by speaking rates. Additionally, we divide the training and validation sets, such that only unseen utterances are used for validation. During this stage, we use a batch size of 5, the validation TV RMSE as our model selection metric, set $\lambda = 0.4$, $N = 60$ with shorter phoneme sequences being padded, and $f_c = 10$, resulting in a 10 Hz low-pass filtering. Finally, the implementation of the FS loss was taken from [26], and the low-pass layer from [27].

At inference, the trained `f-APTAI` model produces three distinct outputs for a given input audio file: nine TV trajectories, a phoneme sequence, and a phoneme alignment. We evaluate the first via PCC and RMSE between the predicted and ground truth HPRC-N trajectories. For phoneme-related evaluation, we use PER for the phoneme sequence, where the ground truth and upper bound of our method are phoneme labels created by the WebMAUS forced aligner, which is also used to measure the overlap metric.

5.2 APTAI Pathological Performance

We use the best-performing model trained during the first experiment (see Sect. 5.1) to perform inference on TORGO, which contains EMA sensor data for dysarthric speech. This experiment aims to see how well the `f-APTAI` model can predict with pathological speech as input. To this end, we selected one patient from each severity group with the required data (mild, moderate, severe) and two control speakers (male and female). It should be noted, that all severe patients in this corpus have some EMA sensor data missing. We chose the severe patient “M02" because he had the least data missing. However, we still had to remove the “TBCL" TV from all speakers since this patient’s EMA data was corrupted and we wanted to keep the results comparable between the groups. Overall, this enabled us to predict 5/9 TVs since three other TVs had to be removed because of missing palate data (see Sect. 4.1). For each selected speaker, we only consider files from one recording session. Furthermore, we selected the head-microphone audio files as inference basis, as they were less noisy (but still noisy), than the ones recorded from the array-microphones. Lastly, we trimmed silence from the beginning and end of the recordings.

5.3 Dysarthria Severity Classification

This experiment assesses whether the trained and frozen `f-APTAI` model can extract meaningful hidden features for downstream tasks. We apply this to dysarthria severity classification using the UASpeech dataset. Figure 3 illustrates this use case, showing both four- and six-class classifications, depending on whether “TTS" and “Control" classes are included. We select a balanced number of UASpeech speakers per class, with three speakers per group and 448 utterances per speaker. Table 2 lists the selected UASpeech speakers per group.

We experiment with two different setups, where “Setup 1" tests on an unseen speaker, and “Setup 2" tests on a speaker that was seen during training (but unseen utterances). In the seen speaker “Setup 2", data (balanced per class) is randomly split 70/15/15 for training, validation, and testing. We report the mean and standard deviation across three training runs (i.e., random splits). For “Setup 1", we ran three tests, each with a different speaker excluded from training. The remaining two speakers per class are split 70/30 for training and validation.

Table 2 Selected speakers per UASpeech intelligibility group for the conducted experiments

Class	Speakers	Intelligibility [%] ↑
TTS	TM01, TF01, TF02	–
Control	CM01, CM04, CF03	–
High	M10, F05, M14	93, 95, 90
Mid	M05, M11, F04	58, 62, 62
Low	M07, F02, M16	28, 29, 43
Very low	M01, M04, F03	15, 2, 6

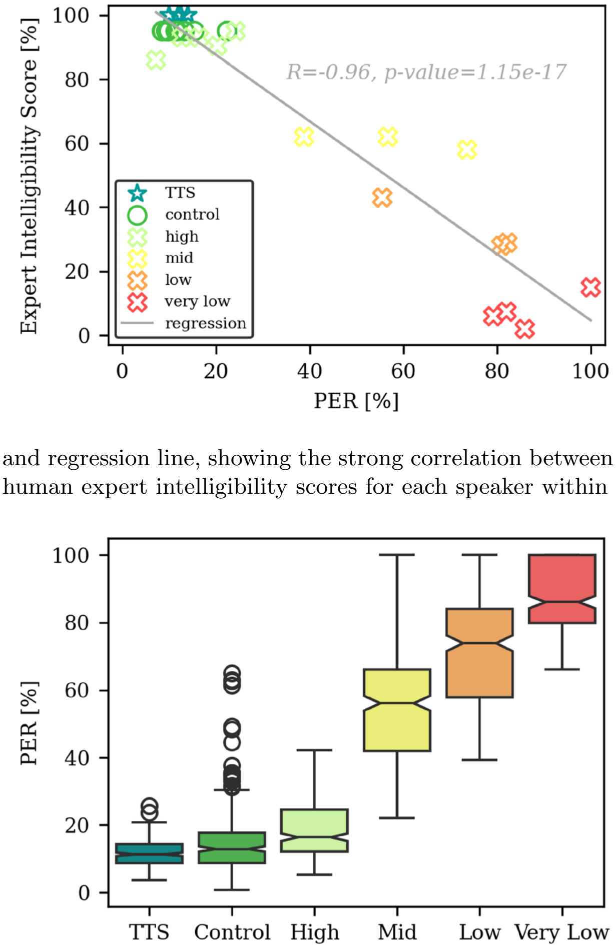
For hidden feature extraction, we select the best-performing `f-APTAI` model trained during the first experiment (see Sect. 5.1). We chose 32 epochs, a batch size of 1, a learning range of 1e−4, a linear learning rate scheduler, and set dropout values (in order of occurrence) to 20, 10, 10%. Lastly, model performance is evaluated in terms of classification metrics F1-score and recall, and model selection is based on the validation F1-score.

5.4 Intelligibility Estimation

This experiment examines Step I of the proposed framework (see Fig. 2 and Sect. 3.1). We analyze how `f-APTAI`-based PER correlates with human expert intelligibility scores. We use UASpeech recordings from all speakers per intelligibility group. We conduct two experiments, varying N utterances per speaker across k runs. To analyze PER correlation with human scores (for individual speakers and groups), we set $N = 448$ (maximum value) and $k = 1$. Additionally, we test a more practical scenario ($N = 25$, e.g., in therapy) and analyze PER variation across groups over $k = 100$ runs (randomly selected utterances). PER may exceed 100%, particularly for “Very Low" intelligibility patients when numerous insertions are required. In such cases, we cap PER at 100%. For visualization (Fig. 4), we assume intelligibility scores of 100% for TTS utterances and 95% for control speakers, as UASpeech lacks labels for them.

5.5 Analysis of Phonetic Groups

This experiment evaluates the validity of Step II in our framework (see Fig. 2 and Sect. 3.2). We use UASpeech speakers, including two randomly selected from the “TTS" and “Control" groups. We compute the recognition rate for IPA-based phonetic groups (see Table 1) using all 448 utterances per speaker. This analysis relies on `f-APTAI`’s predicted phoneme sequence and the grapheme-to-phoneme conversion of utterance transcriptions. We visualize results using a heat map for better interpretability.



(a) Scatter plot and regression line, showing the strong correlation between mean PER ($N = 448, k = 1$) and human expert intelligibility scores for each speaker within the groups.

(b) Spread of the mean PER for the individual speaker groups when using $N = 25, k = 100$.

Fig. 4 Intelligibility analysis (framework Step I), based on the mean PER between a grapheme-to-phoneme conversion and the predicted `f-APTAI` phoneme sequence. Considering N random utterances over k runs (UASpeech dataset)

5.6 Using TTS as Reference

This experiment evaluates the practicality of using a TTS reference in Step III (see Fig. 2 and Sect. 3.3) of our framework. We conduct four sub-experiments to gather relevant insights. In the main experiment, we generate TTS versions of HPRC-N utterances and use the best-performing `f-APTAI` model (see Sect. 5.1) to predict TV trajectories for both. Next, we apply dynamic time warping (DTW) [38] (*librosa* implementation) to align time resolutions. This ensures that the TTS version's correlation with the original (PCC and RMSE metrics) is measured, determining whether it can replace the original without information loss.

Additionally, we conduct a similar experiment using a subset of the UASpeech dataset (see Table 2) to examine differences between TTS-generated utterances and dysarthric severity groups, as UASpeech lacks EMA sensor data. To address this, we use the `f-APTAI` model to predict TV trajectories for all recordings. As in the TORGO experiments, we remove initial and trailing silence from the UASpeech audio files. We then compute the mean RMSE between TTS and different UASpeech speakers, applying the DTW algorithm to align time axes. The results are visualized as a radar plot, providing an interpretable comparison.

Another experiment visualizes TTS and UASpeech utterance pairs across dysarthric severity levels. We visualize the full `f-APTAI` model output (nine TVs, phoneme sequence, and alignment) to demonstrate Step III's application, e.g., in therapy. Since the UASpeech corpus lacks EMA sensor-related labels, further metric computation is not possible.

Finally, we include TTS and control speakers in Experiment 5.4 to compare their intelligibility estimation via `f-APTAI` predicted phoneme sequences and resulting PER.

5.7 Phoneme-Based Keyword Detection

This experiment evaluates the effectiveness of a phoneme sequence-based keyword detection approach (open vocabulary), proposed as Step IV (see Fig. 2 and Sect. 3.4) of our framework. Since this method is highly sensitive to phoneme prediction errors, we introduce a tolerance level τ (absolute phoneme count) in our experiments. This allows keyword detection even when some phonemes do not match exactly. We conduct this experiment using the UASpeech dataset, including all speakers and groups, as well as the "TTS" and "Control" categories. Each file contains a single utterance, for which we predict phoneme sequences using the `f-APTAI` model and generate grapheme-to-phoneme conversions via the WebMAUS-API (based on dataset transcripts). Finally, we determine whether the predicted phoneme sequence aligns with the converted target keyword within the tolerance level τ, evaluating performance in terms of keyword detection accuracy.

6 Results and Discussion

This section presents and discusses the results of the seven experiments described in Sect. 5, maintaining the same chronological order. Overall, the experiments validate the proposed framework (Sect. 3) while also providing additional insights into how linguistic structures, such as phonemes and their articulatory realizations, can be effectively modeled using deep learning. The results demonstrate that the integration of phoneme recognition, phoneme alignment, and articulatory trajectory prediction bridges the gap between abstract linguistic units and their concrete articulatory realizations, reinforcing the role of EMA-based datasets in capturing speech dynamics.

6.1 APTAI Performance

This section presents the performance results of the `f-APTAI` model for the novel APTAI task, introduced in our prior work. Figure 1b illustrates model predictions compared to ground truth values. The main results from the LOSO test setup are shown in Table 3. The model achieves an average PCC of 0.71 and RMSE of 0.68 mm, with minimal variation across the eight test speakers, demonstrating speaker independence. This performance highlights the utility of

Table 3 Test results (HPRC) for the `f-APTAI` model in terms of prediction metrics, for unseen test speakers. The first and second rows per speaker correspond to "normal" and "fast" speaking rates (i.e., HPRC-N and HPRC-F subsets), with bold values indicating the respective averages

Test	TV metrics		Phoneme metrics	
Speaker	PCC↑	RSME [mm]↓	PER [%]↓	Overlap [%]↑
M01	0.69 ± 0.22	0.69 ± 0.21	3.64 ± 4.0	76.71 ± 5.3
	0.63 ± 0.19	0.76 ± 0.16	11.60 ± 9.9	72.20 ± 7.2
M02	0.65 ± 0.21	0.74 ± 0.19	4.24 ± 4.5	76.57 ± 5.5
	0.61 ± 0.21	0.78 ± 0.18	15.72 ± 11.8	69.80 ± 8.3
M03	0.72 ± 0.11	0.68 ± 0.15	3.80 ± 4.0	78.36 ± 4.9
	0.67 ± 0.15	0.73 ± 0.14	9.57 ± 8.6	74.53 ± 7.1
M04	0.71 ± 0.11	0.69 ± 0.12	3.96 ± 3.9	76.68 ± 5.4
	0.66 ± 0.09	0.74 ± 0.09	6.26 ± 6.4	75.67 ± 6.3
F01	0.71 ± 0.22	0.68 ± 0.22	3.92 ± 4.0	76.37 ± 5.1
	0.63 ± 0.20	0.77 ± 0.18	6.81 ± 6.9	74.90 ± 6.5
F02	0.73 ± 0.14	0.67 ± 0.15	6.12 ± 5.3	72.46 ± 6.3
	0.67 ± 0.13	0.73 ± 0.13	13.63 ± 10.8	68.24 ± 8.8
F03	0.72 ± 0.14	0.68 ± 0.15	4.17 ± 4.2	76.78 ± 5.7
	0.64 ± 0.15	0.75 ± 0.14	5.39 ± 6.0	77.06 ± 5.9
F04	0.75 ± 0.16	0.63 ± 0.17	5.06 ± 5.1	75.60 ± 5.5
	0.71 ± 0.14	0.68 ± 0.14	13.35 ± 11.1	71.03 ± 8.3
Avg.	**0.71 ± 0.03**	**0.68 ± 0.03**	**4.36 ± 0.8**	**76.19 ± 1.7**
	0.65 ± 0.04	0.74 ± 0.03	10.29 ± 3.9	72.93 ± 3.1

EMA technology, as its high spatiotemporal resolution enables precise articulatory tracking, which is particularly beneficial for phoneme-to-articulatory mappings. Compared to alternative methods such as ultrasound or MRI, EMA facilitates dynamic articulatory motion capture, making it a compelling choice for deep learning-based speech analysis [39]. Phoneme-related metrics indicate strong performance, with a mean PER of 4.36% and a 76.19% overlap with a SOTA text-dependent forced aligner. Table 4 presents TV metrics for HPRC-N and HPRC-F (normal and fast speaking rates), showing that rear tongue constriction degree predictions are particularly error-prone. A similar effect was observed in our prior work with the base `APTAI` model. Despite targeted investigations, the lower performance in rear tongue constriction predictions remains unresolved, as no clear technical cause was identified. We hypothesize that this may be due to phonetic context, where other tract variables provide more distinctive articulatory cues, making rear tongue constriction inherently more difficult to predict. As expected, faster speaking rates lead to reduced prediction accuracy due to the increased complexity of the task. While improvements are always possible, the model is sufficiently robust to serve as the cornerstone of the proposed speech and language analysis framework.

6.2 APTAI Pathological Performance

This experiment evaluated `f-APTAI`'s prediction capabilities on a new corpus (TORGO), which includes pathological (dysarthric) speech. Table 5 presents TV prediction results (PCC, RMSE). Overall, performance declines compared to the HPRC dataset (Table 3), with dysarthric speakers exhibiting the largest drop. The two control speakers perform slightly worse than HPRC-N speakers, with a PCC decrease of less than 0.1 and similar RMSE values. Performance consistently deteriorates as dysarthria severity increases, indicating that the APTAI task becomes more challenging as dysarthric effects intensify. These results align with Experiment 5.6, visualized as a radar plot (Fig. 5).

Table 4 Individual TV metrics, in terms of mean and deviation across the `f-APTAI` leave-one-speaker-out experiments

TV's	HPRC-N		HPRC-F	
	PCC↑	RSME [mm]↓	PCC↑	RSME [mm]↓
LA	0.85 ± 0.03	0.53 ± 0.04	77.03 ± 4.7	0.63 ± 0.06
LP	0.76 ± 0.07	0.64 ± 0.08	66.75 ± 5.7	0.73 ± 0.05
JA	0.81 ± 0.03	0.58 ± 0.04	70.59 ± 4.5	0.70 ± 0.05
TTCL	0.81 ± 0.04	0.58 ± 0.06	76.74 ± 4.2	0.63 ± 0.05
TTCD	0.78 ± 0.05	0.63 ± 0.06	69.71 ± 4.3	0.71 ± 0.04
TMCL	0.79 ± 0.03	0.60 ± 0.04	74.63 ± 3.5	0.65 ± 0.04
TMCD	0.34 ± 0.10	1.05 ± 0.08	29.72 ± 12.5	1.07 ± 0.09
TBCL	0.75 ± 0.04	0.65 ± 0.05	71.49 ± 3.4	0.69 ± 0.04
TBCD	0.51 ± 0.12	0.88 ± 0.10	50.30 ± 12.3	0.87 ± 0.10

Table 5 TV metrics of `f-APTAI` model predictions for selected TORGO recordings and dysarthria severity levels. Only considering 5/9 TVs since some EMA data is missing

Speaker	Severity	#Utterances	PCC↑	RMSE [mm] ↓
FC03	Control	392	0.64 ± 0.23	0.80 ± 0.17
MC01	Control	423	0.62 ± 0.22	0.82 ± 0.16
F04	Mild	252	0.46 ± 0.24	0.93 ± 0.16
F03	Moderate	216	0.37 ± 0.25	0.95 ± 0.19
M02	Severe	169	0.19 ± 0.25	1.14 ± 0.20

TORGO has previously been identified as a noisy dataset [40], with both audio and EMA data suffering from quality issues, including problems with EMA coils in nearly all pathological sessions. To mitigate these issues, we limited predictions to 5/9 TVs from selected speakers and sessions. Despite these challenges, certain observations can be made. The clear correlation between declining performance and dysarthria severity suggests that severely affected patients may benefit from first improving overall intelligibility (Step I, Sect. 3.1) before undergoing detailed articulatory analysis (Step III, Sect. 3.3). This may stem from the model's reliance on hidden acoustic phoneme information, which likely requires a baseline level of intelligibility for accurate TV predictions.

While our approach prioritizes training on typical speech to ensure generalization and avoid dependence on limited pathological datasets, an alternative strategy could involve incorporating synthetically degraded speech. However, generating clinically meaningful synthetic data poses

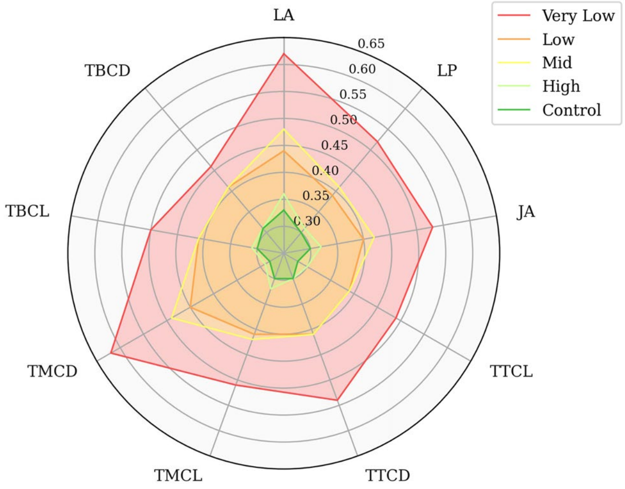


Fig. 5 Mean RMSE [mm] ↓ prediction results between TTS and UASpeech intelligibility groups. "High" and "Control", as well as "Mid" and "Low" results are similar

challenges, requiring either well-labeled pathological datasets for generative models or transformation techniques that realistically modify healthy speech. Both approaches must balance realism, avoid bias, and ensure broad applicability. Future work could explore this direction while addressing these limitations.

6.3 Dysarthria Severity Classification

This experiment aimed to determine whether embeddings extracted from the trained and frozen `f-APTAI` model could be used for downstream tasks, specifically dysarthria severity estimation. Dysarthria is associated with impaired articulator control, often affecting phoneme realization, which can lead to inconsistent or distorted articulatory movements. By modeling phoneme sequences and their temporal alignment alongside articulatory trajectories, our approach explicitly captures the transition from abstract phonemic representations to their concrete articulatory executions. This structured mapping allows for a more interpretable analysis of dysarthria severity levels, reinforcing the model's role in clinical applications. Our model's phoneme sequence and phoneme alignment predictions provide a foundation for assessing severity levels. Since phoneme realization is directly linked to intelligibility, the learned embeddings may capture articulatory distortions and compensatory strategies that reflect severity differences. To maintain interpretability, we deliberately employed a basic architecture with approximately 300, 000 learnable parameters. This task is inherently challenging because expert severity labels are assigned at the speaker level, whereas classification is performed at the utterance level. Not all 448 utterances per speaker necessarily fit the same severity category, as shown in Experiment 5.4 and Fig. 4b, where substantial variability in PER is observed across 100 random utterance selections per speaker.

Table 6 presents classification results, while Fig. 6 visualizes confusion matrices for both the four- and six-class experiments. As expected, classification is more difficult in the unseen speaker setup (Setup 1) compared to Setup 2, where the speaker (but not the utterance) was seen during training. Interestingly, performance improves when using six severity classes rather than four, possibly because the TTS speaker is easily identifiable.

Table 6 Results of the dysarthric severity classification experiment of UASpeech speakers (`f-APTAI` hidden features)

Group	Setup 1		Setup 2	
	Recall [%]↑	F1 [%]↑	Recall [%]↑	F1 [%]↑
Four-class classification				
High	79.3 ± 14.4	74.3 ± 4.5	79.3 ± 2.5	81.3 ± 1.2
Mid	11.0 ± 6.6	11.7 ± 7.1	64.3 ± 6.5	62.7 ± 4.0
Low	25.0 ± 20.5	23.0 ± 16.5	65.7 ± 4.6	69.0 ± 1.7
Very Low	45.0 ± 24.0	44.3 ± 21.8	80.7 ± 3.1	76.7 ± 0.6
Avg.	40.1 ± 29.6	38.3 ± 27.5	72.5 ± 8.7	72.4 ± 8.2
Six-class classification				
TTS	98.3 ± 1.2	99.3 ± 1.2	100 ± 0.0	100 ± 0.0
Control	62.3 ± 21.1	66.0 ± 14.8	86.7 ± 1.5	86.7 ± 0.6
High	55.7 ± 16.6	56.0 ± 8.7	77.7 ± 6.7	76.7 ± 3.8
Mid	15.0 ± 9.6	13.0 ± 7.9	67.3 ± 6.7	62.3 ± 0.6
Low	16.7 ± 9.7	16.0 ± 7.2	69.3 ± 3.1	70.3 ± 2.5
Very Low	43.7 ± 26.9	45.3 ± 22.5	71.0 ± 5.6	75.6 ± 2.5
Avg.	48.6 ± 31.2	49.3 ± 32.5	78.7 ± 12.6	78.6 ± 13.2

The “Mid" severity class proves the most challenging to classify, likely due to inconsistencies between speaker-level severity labels and utterance-level articulatory characteristics. Dysarthria severity can fluctuate across speech samples, making single-category speaker labels an imperfect reflection of utterance-level impairments. This issue is further reinforced by Experiment 5.4 (Fig. 4b), where PER deviations across utterances indicate inconsistencies in severity representation. Given the scarcity of publicly available pathological datasets, we rely on UASpeech despite its speaker-level severity annotations. Future work

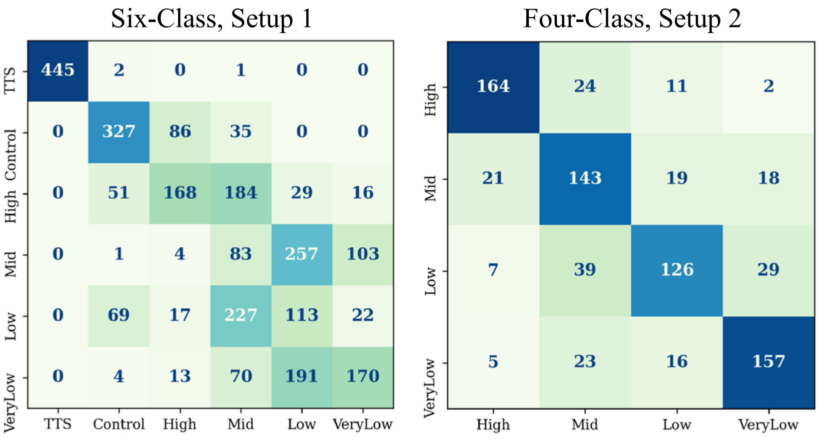


Fig. 6 Confusion matrices of the hidden feature extraction experiment for dysarthria severity classification

could explore dataset curation strategies to provide more fine-grained, utterance-specific severity labels and improve classification performance.

Comparing performance for this classification task is inherently difficult. To the best of our knowledge, SOTA speaker-independent four-class classification performance [41, 42] ranges from an F1-score of 40 to 49%, whereas our model achieves 38.3% in this scenario. However, our setup is not directly comparable to SOTA benchmarks, as our primary goal was not to achieve state-of-the-art performance but rather to evaluate whether our novel model can support hidden feature extraction for downstream tasks in a deliberately simple setup.

Since `f-APTAI` provides a generalizable embedding space, the optimal architecture for utilizing these embeddings should be task-specific, as different applications require tailored network designs. Future work could explore specialized architectures optimized for individual downstream tasks. Additionally, while our current approach relies on phoneme embeddings, future research could explore phoneme-group embeddings to capture systematic pronunciation variations across phoneme classes. Such an approach could enhance robustness in pathological speech modeling but would require a fundamental redesign of the model architecture.

6.4 Intelligibility Estimation

This experiment investigated whether the predicted phoneme sequence from the trained and frozen `f-APTAI` model, combined with a grapheme-to-phoneme conversion of target utterances, can serve as a speaker-level intelligibility estimator. The main results are presented in Fig. 4 as a scatter plot (Fig. 4a) and box plot (Fig. 4b). The scatter plot illustrates the correlation analysis between PER and human expert intelligibility labels, revealing a strong correlation. The box plot visualizes the PER distribution for $N = 25$ utterances over $k = 100$ random selections from a total of 448 utterances per speaker. These results suggest that 25 utterances are sufficient for real-world intelligibility estimation. The plot also reinforces previous findings that the “Mid" severity class is particularly challenging, as evidenced by the relatively large spread in PER values. Additionally, a high number of outliers in the “Control" group contrasts with only two outliers in the “TTS" group (artificial speakers). This finding should encourage the use of artificially generated speech instead of control groups for experiments, providing more consistency. Overall, these results validate Step I of the proposed framework, demonstrating its potential for improving speech intelligibility. Since Steps III and IV rely on intelligible input, Step I is a crucial prerequisite—without it, the later steps could be prone to errors, as observed in other experiments.

6.5 Analysis of Phonetic Groups

This experiment extends Experiment 5.4 by investigating the recognition rate of specific phonetic groups. The results are visualized in Fig. 7, with Table 1 detailing the relationship between individual phonemes and phonetic groups. Since this experiment uses the UASpeech corpus, which contains dysarthric speech, the findings align with expectations: dysarthria affects articulation globally rather than specific phonetic groups. A closer look at “Mid" severity level speakers “M05" and “M11", who have similar expert intelligibility ratings, reveals differences in phonetic group recognition. Speaker ”M05" shows particular difficulty with “postalveolar" and “trill" production, whereas for “M11", all groups are affected similarly. Applying these findings to our proposed framework, “M05" presents an “identified problem", where therapy could focus on utterances and exercises targeting these specific phonetic groups.

6.6 Using TTS as Reference

Step III of the proposed speech and spoken language analysis framework leverages an artificially generated reference for target utterances. This approach is motivated by generative AI techniques, which learn from diverse datasets to produce idealized “average speakers”. This experiment evaluates the validity of using TTS-generated speech as a reference, a concept that could be extended to other applications. Figure 8 illustrates an example of Step III, where a TTS reference is analyzed alongside a pathological utterance. In Fig. 8a, a dysarthric speaker with “High” intelligibility produces an utterance closely matching the TTS reference, with only a minor time difference. In contrast, Fig. 8b highlights a specific articulation error in a “Mid" intelligibility speaker’s pronunciation of “python", where the target fricative /θ/ is replaced by the fricative /s/. This substitution is also visible in the mel spectrogram, showing increased high-frequency energy during the relevant time frames. The model predictions confirm this misarticulation, as seen in phoneme sequence, alignment, and articulatory feature outputs, particularly in TMCD (tongue middle constriction degree), which shifts noticeably during the articulation of /θ/. This visualization of the model’s predictions also highlights the physical effort required for articulation in this dysarthric patient. The impact of dysarthria is evident in the TV trajectories and along the time axis, where vowel productions (/aɪ/ and /ɒ/) are particularly affected. Figure 8c further demonstrates that Step III is ineffective for patients with “Very Low” intelligibility, where the articulation deviations are too severe for meaningful analysis. In such cases, iterating Step I (speech intelligibility improvement) is far more beneficial before advancing to more detailed articulatory assessments.

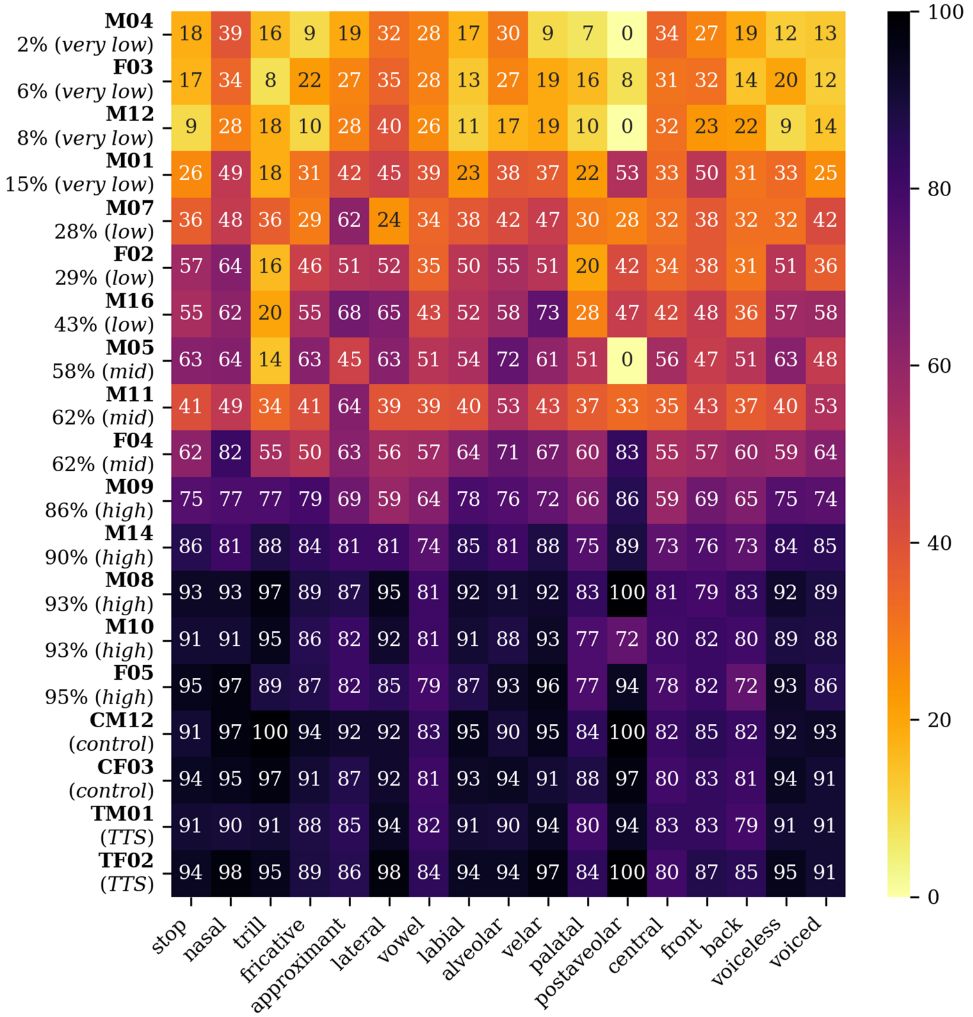


Fig. 7 Heatmap of UASpeech and TTS speakers, showing the recognition rate of phoneme groups (framework Step II/III), considering all 448 utterances per speaker

Beyond individual examples, the main investigation results are presented in Fig. 9 and Table 7. The figure illustrates time discrepancies between TTS and human utterances, which we correct using Dynamic Time Warping (DTW). The correlation analysis in Table 7 confirms a strong speaker-independent correlation (around 0.92) between predicted TTS-based TVs and ground truth HPRC-N human TVs, validating our approach of using TTS-generated speech as a reference. Additional evidence from Fig. 4b and Table 6 further supports the effectiveness of TTS-based references within our framework. A similar approach also holds potential in a therapeutic setting: when designing specific exercises, the TTS-generated versions could be automatically created, providing more consistency and flexibility over control speakers.

Finally, an additional experiment explored differences in model predictions (TVs) between TTS and pathological UASpeech utterances. The results, visualized as a radar plot (Fig. 5), show consistent deviations in mean RMSE across dysarthria severity groups, aligning with human expert severity ratings. Three distinct severity clusters emerge. ”High” and ”Control” speakers are closest to TTS, as expected. ”Mid” and ”Low” severity speakers exhibit higher deviations from TTS utterances, while ”Very Low” severity cases show the largest deviation, indicating severe articulatory disruption relative to typical speech. These findings align with prior results. Figure 8 provides visual articulation examples, while Fig. 4 confirms intelligibility relations. The dysarthria severity classification results in Table 6 reinforce that the ”Mid" severity class remains the most challenging to classify and analyze.

6.7 Phoneme-Based Keyword Detection

This experiment aimed to determine whether the predicted phoneme sequence of the `f-APTAI` model could be effectively used for open vocabulary keyword detection, particularly in applications such as aphasia therapy. This is especially relevant for stroke survivors with aphasia, where impaired word retrieval can provide valuable diagnostic

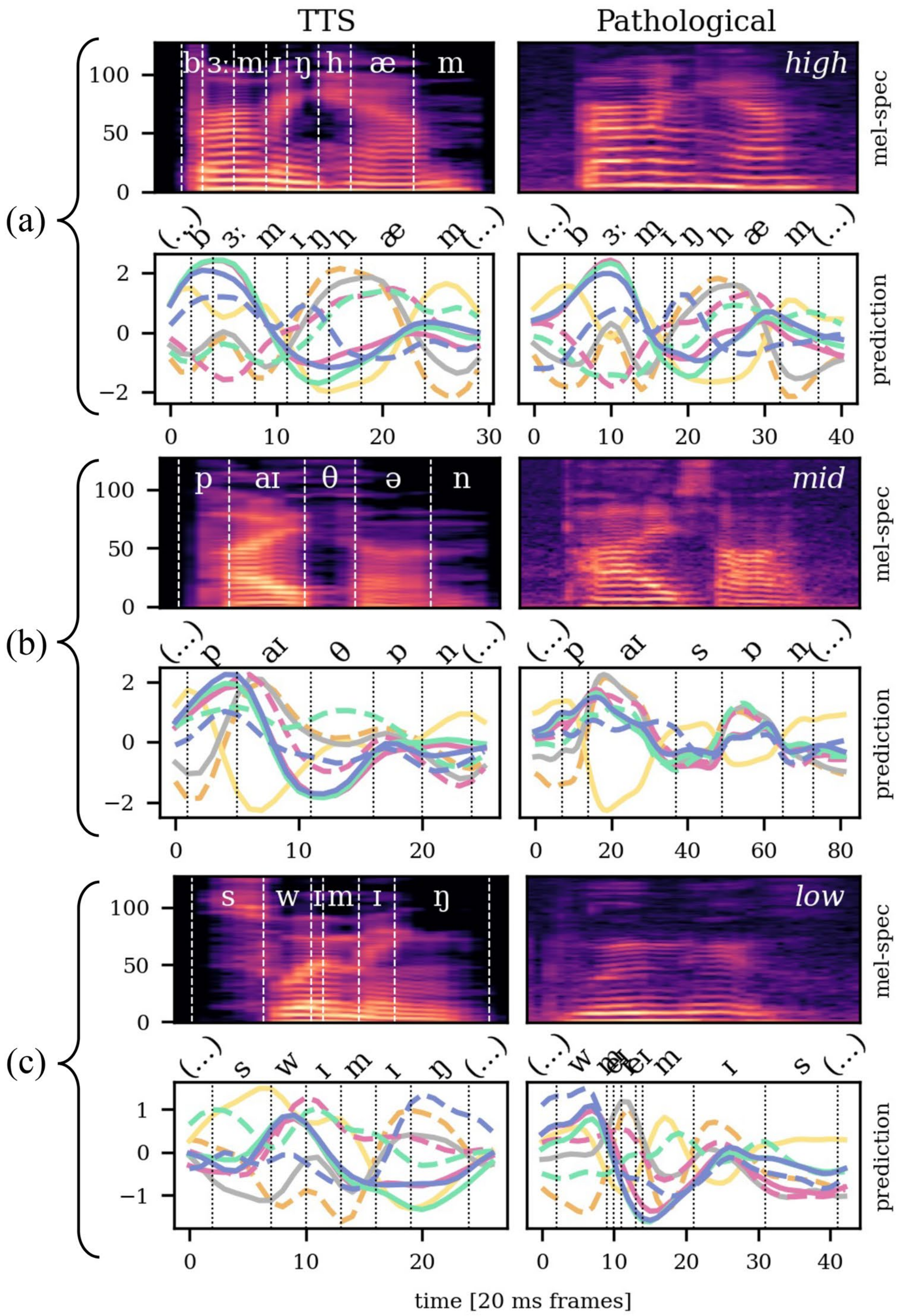


Fig. 8 Comparison between three utterances ("Birmingham", "python", and "swimming") of a TTS generated and a pathological (UASpeech) speaker with different expert intelligibility labels ("high", "mid", "low") in terms of `f-APTAI` model prediction (framework Step IV)

insights. Figure 8a illustrates this concept with an example of the utterance "Birmingham", which would have been correctly detected in this case.

The main results are presented in Table 8, evaluated at different phoneme-based tolerance levels (τ). Examining this table alongside Fig. 4, a clear connection emerges between PER and keyword detection accuracy. Choosing an appropriate tolerance level requires careful consideration, as any change in a phoneme inherently alters the meaning of a word. However, a tolerance level of 1 may be acceptable depending on the application, as it enhances robustness in keyword detection. For example, even the smallest possible tolerance level might be problematic in the context of aphasia, dependent on the chosen test utterances. In contrast, Klumpp [43] recently introduced a more robust German keyword detection method

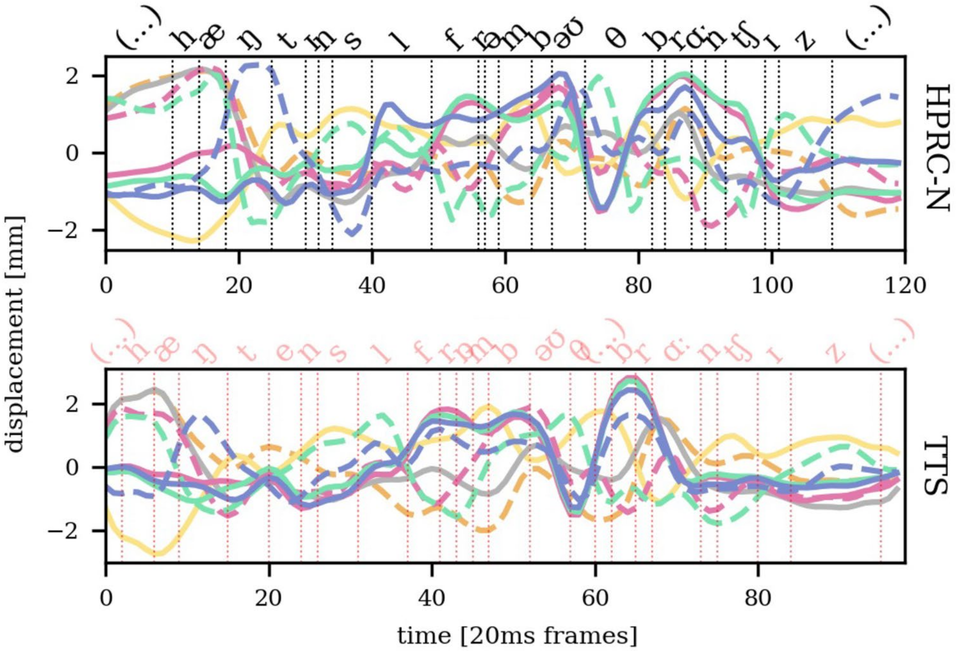


Fig. 9 `f-APTAI` prediction of an HPRC-N and TTS generated utterance (speakers: M03 and TM01): “hang tinsel from both branches", showcasing the time difference

Table 7 Mean and deviation of PCC↑ (first row) and RMSE [mm] ↓ (second row) across four male (M) and female (F) HPRC-N speakers and four aligned TTS speakers

Gender	TM01	TM02	TF01	TF02
M (avg.)	0.93 ± 0.01	0.92 ± 0.07	0.92 ± 0.06	0.93 ± 0.05
	0.37 ± 0.03	0.37 ± 0.02	0.38 ± 0.01	0.36 ± 0.01
F (avg.)	0.92 ± 0.07	0.92 ± 0.02	0.92 ± 0.04	0.93 ± 0.06
	0.38 ± 0.01	0.38 ± 0.01	0.38 ± 0.01	0.37 ± 0.02

in a pathological context. This approach improves reliability by incorporating phonemes from other languages and using a transformer-based architecture to capture broader pronunciation variations of the same keyword. Therefore, this method is robust to pronunciation errors, which might be caused by a speech disorder but should not affect a cognitive ability analysis (e.g., aphasic context).

Future work could explore integrating a similar phoneme-variant retrieval module into the `f-APTAI` model. However, this would require EMA sensor articulation data from multiple languages, which remains a significant challenge due to the scarcity of cross-lingual EMA datasets. Given this limitation, we chose not to pursue this direction in the current study. If suitable multilingual datasets become available, future research could investigate this approach further.

7 Summary and Conclusion

This article presented a speech and spoken language analysis framework centered around an E2E deep-learning model (`f-APTAI`). Unlike traditional ensemble-based methods that require managing distinct datasets and retraining separate models, our E2E approach leverages inherent output relationships, simplifying the analytical pipeline and enhancing robustness. This study marks progress toward a fully

Table 8 Mean and deviation (framework step V) of keyword detection accuracy ↑ [%] across UASpeech speaker groups, with a tolerance level τ (tolerated mismatched phonemes). Based on the phoneme sequence prediction

Tolerance	TTS	Control	High	Mid	Low	Very low
$\tau = 0$	67.8	60.9	50.6	14.3	5.3	0.3
	± 3.0	± 8.5	± 9.4	± 9.3	± 5.8	± 0.5
$\tau = 1$	83.5	75.9	65.0	24.0	10.0	2.0
	± 1.7	± 8.2	± 9.9	± 11.5	± 9.5	± 0.8
$\tau = 2$	88.5	83.9	73.8	34.0	17.0	6.3
	± 2.4	± 5.6	± 9.2	± 11.4	± 9.6	± 2.6
$\tau = 3$	91.8	88.7	80.6	46.3	28.7	18.3
	± 1.9	± 4.5	± 6.4	± 11.0	± 8.6	± 3.4

integrated framework for multi-faceted speech and language analysis.

For a given audio input, our model predicts speaker- and text-independent articulation trajectories, paired with the corresponding phoneme sequence and alignment, enabling analysis at various levels of abstraction. The framework has various potential applications, such as (self-)therapy via a smartphone app. However, any automated framework intended for therapeutic use must be highly reliable and rigorously evaluated in its specific environment. While our `f-APTAI` model performs competitively, further improvements are needed before it or a similar model can be used in this context. Considering this, we presented a blueprint for developing similar frameworks with a robust AI model at their core.

Experiments 5.1 and 5.2 assessed general and pathological `f-APTAI` performance, with competitive results in the former and promising outcomes in the latter. Experiment 5.3 explored whether the model's hidden representations could serve as features for downstream tasks, using dysarthria severity classification as an example. Experiments 5.4 and 5.5 examined Steps I and II of our framework, confirming that PER can serve as a biomarker for speech intelligibility and that phoneme recognition rates can help identify articulation problems, a requirement for Step III. Experiment 5.6 demonstrated the viability of TTS as a supporting reference, enabling detailed problem analysis. Lastly, Experiment 5.7 evaluated Step IV, showing that open-vocabulary keyword detection remains an area for improvement.

In conclusion, this work underscores the promise of an E2E deep learning approach in bridging the gaps of traditional methods, paving the way for scalable and adaptable solutions in speech and spoken language processing. Future work will expand model capabilities by incorporating fundamental frequency estimation. Additionally, adapting recent findings on robust open-vocabulary keyword detection could enhance analysis. Another research direction is improving automatic speech generation by integrating `f-APTAI`'s hidden articulation and alignment embeddings. Further research will also explore phoneme-group embeddings to capture systematic pronunciation variations and apply the framework to additional pathological datasets, such as cleft lip and palate.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44230-025-00094-6.

Author Contributions Tobias Weise contributed to the conceptualization, software, data curation, methodology, formal analysis and investigation, and writing-original draft preparation. Kubilay Can Demir contributed to the formal analysis and investigation, writing-review and editing. Paula Andrea Pérez-Toro contributed to the formal analysis writing-review and editing. Tomas Arias-Vergara contributed to the formal analysis writing-review and editing. Andreas Maier contributed to the writing-review and supervision. Elmar Nb̈oth contributed to the writing-review and supervision. Maria Schuster contributed to the writing-review and editing and supervision. Björn Heismann contributed to the writing-review and supervision. Seung Hee Yang contributed to the writing-review and supervision.

Funding We gratefully acknowledge funding for this study by Friedrich-Alexander-University Erlangen-Nürnberg, Medical Valley e.V. and Siemens Healthineers AG within the framework of d.hip campus.

Data Availability For additional information about the corpora used in this study, please refer to https://experts.illinois.edu/en/publications/dysarthric-speech-database-for-universal-access-research for the UASpeech, https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html for TORGO, and https://huggingface.co/pklumpp/Wav2Vec2_CommonPhone for Common Phone datasets.

Declarations

Conflict of interest The authors declare they have no conflict of interest.

References

1. Chartier J, Anumanchipalli GK, Johnson K, et al. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron. 2018;98(5):1042–54.
2. Narayanan SS, Nayak KS, Lee S, Sethy A, Byrd D. Magnetic resonance imaging and electromagnetic articulography database for speech research. J Acoust Soc Am. 2004;121(5):3074–82.
3. Tanabe H, Ertan A, Byrd D, Narayanan SS. Vocal tract imaging using real-time MRI for articulatory phonetics and speech science research. Phon Speech Sci. 2021;13(1):47–59.
4. Bernhardt BM, Gick B, Bacsfalvi P, Adler-Bock M. Ultrasound in speech therapy with adolescents and adults. Clin Linguist Phon. 2005;19(6–7):605–17. https://doi.org/10.1080/02699200500113943.
5. Bradlow AR, Lee E-K. The use of MRI and ultrasound technology in teaching about Spanish and general phonetics and pronunciation. In: Proceedings of the 6th International Conference on Spanish Phonetics and Phonology, 2015.
6. Browman CP, Goldstein L. Gestural specification using dynamically-defined articulatory structures. J Phon. 1990;18(3):299–320.
7. Ji A. Speaker independent acoustic-to-articulatory inversion. PhD thesis, Marquette University, 2014.
8. McGowan RS. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model tests. Speech Commun. 1994;14(1):19–48.
9. Weise T, et al. Speaker-and text-independent estimation of articulatory movements and phoneme alignments from speech. 2024. arXiv preprint (INTERSPEECH) arXiv:2407.03132.

10. Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods. 2018;151:41–54.
11. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alzheimer's Res Therapy. 2022;14(1):186.
12. Mohamed A-R, Dahl GE, Hinton G. Deep belief networks for phoneme recognition. Adv Neural Inf Process Syst. 2012;25:448–56.
13. Wu Z, King S, Renals S. Articulatory-to-acoustic conversion using deep neural networks. Speech Commun. 2016;77:165–77.
14. Rudzicz F. Dysarthria recognition using deep belief networks. Int J Speech Technol. 2012;15:213–27.
15. Schatz T, Mitra V, Feldman N, Goldwater S. Phoneme recognition using articulatory features and deep learning. In: Proceedings of INTERSPEECH 2013.
16. Senior A, Narayanan A. Computational challenges in deep learning for speech recognition. IEEE Trans Audio Speech Lang Process. 2014;22:990–1000.
17. Caruana R. Multitask learning. Mach Learn. 1997;28(1):41–75.
18. Liu H, Chen Y, Yu D. Self-supervised learning for speech representation learning: a survey. IEEE Trans Speech Audio Process. 2019;27:1–10.
19. Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, Liu T-Y. FastSpeech: fast, robust and controllable text to speech. Adv Neural Inf Process Syst. 2019;32:3171–80.
20. Baevski A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst. 2020;33:12449–60.
21. Hendrycks D, Gimpel K. Gaussian error linear units (GELUS). 2016. arXiv preprint arXiv:1606.08415.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.
23. Graves A. Connectionist temporal classification. In: Supervised sequence labelling with recurrent neural networks. Springer; 2012. p. 61–93.
24. Zhu J, Zhang C, Jurgens D. Phone-to-audio alignment without text: a semi-supervised approach. In: ICASSP 2022. IEEE. p. 8167–71.
25. Badlani R, Łańcucki A, Shih KJ, Valle R, Ping W, Catanzaro B. One TTS alignment to rule them all. In: ICASSP 2022, 2022. IEEE. p. 6092–6.
26. Shih KJ, Valle R, et al. Rad-TTS: parallel flow-based TTS with robust alignment learning and diverse synthesis. In: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models, 2021.
27. Parrot M, Millet J, Dunbar E. Independent and automatic evaluation of acoustic-to-articulatory inversion models. In: Proc. Interspeech 2020.
28. Maier A, Haderlein T, Eysholdt U, Rosanowski F, Batliner A, Schuster M, Nöth E. Peaks—a system for the automatic evaluation of voice and speech disorders. Speech Commun. 2009;51(5):425–37.
29. Klumpp P, et al. Common phone: a multilingual dataset for robust acoustic modelling. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, p. 763–8.
30. Ardila R, Branson M, Davis K, Henretty M, Kohler M, Meyer J, et al. Common voice: a massively-multilingual speech corpus. 2019. arXiv preprint arXiv:1912.06670.
31. Garofolo JS. TIMIT acoustic phonetic continuous speech corpus. Linguistic Data Consortium, 1993, 1993.
32. Tiede M, Espy-Wilson CY, et al. Quantifying kinematic aspects of reduction in a contrasting rate production task. J Acoust Soc Am. 2017;141(5– Supplement):3580.
33. Rudzicz F, Namasivayam AK, Wolff T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang Resour Eval. 2012;46:523–41.
34. Wu P, Chen L-W, Cho CJ, Watanabe S, Goldstein L, Black AW, Anumanchipalli GK. Speaker-independent acoustic-to-articulatory speech inversion. 2023. arXiv:2302.06774 [eess.AS].
35. Kim H, Hasegawa-Johnson M, Perlman A, Gunderson J, Watkin K, Frame S. Dysarthric speech database for universal access research. 2008:1741–4
36. Kisler T, Reichel U, Schiel F. Multilingual processing of speech via web services. Comput Speech Lang. 2017;45:326–47.
37. Seneviratne N, Sivaraman G, Espy-Wilson CY. Multi-corpus acoustic-to-articulatory speech inversion. In: Interspeech, 2019, p. 859–63.
38. Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. In: KDD Workshop, vol. 10. Seattle; 1994. p. 359–70.
39. Rebernik T, Jacobi J, Jonkers R, Noiray A, Wieling M. A review of data collection practices using electromagnetic articulography. Lab Phonol. 2021;12(1):1–34. https://doi.org/10.5334/labphon.237.
40. Schu G, et al. On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches. 2022. arXiv preprint arXiv:2211.08833.
41. Joshy AA, Rajan R. Automated dysarthria severity classification: a study on acoustic features and deep learning techniques. IEEE Trans Neural Syst Rehabil Eng. 2022;30:1147–57.
42. Javanmardi F, Kadiri SR, Alku P. Pre-trained models for detection and severity level classification of dysarthria from speech. Speech Commun. 2024;158: 103047.
43. Klumpp P. Phonetic transfer learning from healthy references for the analysis of pathological speech. PhD thesis, Friedrich-Alexander Universität Erlangen-Nürnberg, 2024.