



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/222213/>

Version: Accepted Version

---

**Article:**

Pan, Y., Mirheidari, B., Blackburn, D. et al. (2025) A two-step attention-based feature combination cross-attention system for speech-based dementia detection. *IEEE Transactions on Audio, Speech and Language Processing*, 33. pp. 896-907. ISSN: 1063-6676

<https://doi.org/10.1109/TASLPRO.2025.3533363>

---

© 2025 The Author(s). Except as otherwise noted, this author-accepted version of a journal article published in *IEEE Transactions on Audio, Speech and Language Processing* is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# A Two-Step Attention-based Feature Combination Cross-Attention System for Speech-based Dementia Detection

Yilin Pan<sup>1,2</sup>, Bahman Mirheidari<sup>2</sup>, Daniel Blackburn<sup>3</sup>, and Heidi Christensen<sup>2</sup>, *Member, IEEE*

<sup>1</sup>College of Artificial Intelligence, Dalian Maritime University, China

<sup>2</sup>Department of Computer Science, University of Sheffield, UK

<sup>3</sup>Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, UK

**Abstract**—Dementia poses a significant global challenge, with profound personal, societal, and economic impacts. Although it is incurable, early detection is crucial for ensuring appropriate care and support. Dementia can impair a person’s speech and language abilities, and studies have demonstrated promising results in using spoken language for automatic dementia detection. Recently, deep learning-based self-supervised learning models, such as wav2vec2.0 (w2v) and BERT, have shown success in extracting acoustic and linguistic information. However, most studies have relied on single datasets and relatively straightforward methods for extracting and combining acoustic and linguistic modalities. This paper presents an in-depth exploration of the application of SSL models in this context by proposing the Two-Step Attention-based Feature Combination Cross-attention system (TSAC-ATT) for speech-based dementia detection. The contributions of this paper are as follows: i) we explore and analyse acoustic and linguistic feature extraction pipelines using SSL models, including the proposed TSAC framework to create high-performing acoustic features from w2v’s contextual layers; ii) we demonstrate that these features, when fused using cross-attention, outperform various feature combination approaches; iii) all experimental work is conducted on two publicly available datasets (DementiaBank and ADReSS), as well as the IVA dataset collected by the Royal Hallamshire Hospital, which includes recordings of the standard Cookie Theft task. We present state-of-the-art results, highlighting that acoustic-only features based on the w2v model can achieve very high performance across multiple datasets. Furthermore, we show that the upstream performance of the automatic speech recognition module does not always predict downstream classification performance.

**Index Terms**—Dementia detection, wav2vec2.0, BERT, feature fusion, cross-attention.

## I. INTRODUCTION

**W**ITH an ageing society, the number of people living with dementia is rapidly increasing worldwide. An estimated 55 million individuals were living with dementia

in 2020, and this number is expected to rise to nearly double every 20 years, reaching 139 million by 2050 [1]. The term *dementia* encompasses a range of symptoms associated with the loss of cognitive functioning, including memory, speech and language, visual perception, problem-solving, self-management, attention, and behavioural abilities, all of which can interfere with daily life and activities [2]. The most common cause of dementia is Alzheimer’s Disease (AD).

Common to most causes of dementia, individuals living with AD experience a decline in their speech and language abilities, even in the early stages [3]–[6]. Currently, clinicians use non-invasive manual pen-and-paper assessment tools and invasive diagnostic procedures, such as blood tests [7], to diagnose dementia. However, the accuracy of simpler assessment tools is often unsatisfactory, and further invasive methods, like scans [8], require expert knowledge, are time-consuming, and costly. This places health services under pressure and results in frustratingly long waiting times for patients. Consequently, there is a significant need for automatic, easy-to-use, accurate, and affordable assessments that patients can undertake in clinics or at home to alleviate current bottlenecks and resource demands.

In the past decade, mainstream speech and language processing techniques have seen substantial improvements in performance and robustness, primarily due to increased access to data and advancements in deep learning technologies [9], [10]. Researchers have shown great promise in detecting early signs of cognitive decline that may lead to dementia [11]–[18]. The data used are mostly audio recordings of individuals undertaking various assessments [19], [20]; notably, the picture description task has been the focus of several studies [16], [21]–[23]. The conventional spoken language-based dementia detection system typically consists of a pipeline system that includes a front-end feature extraction module and a back-end classification module [24], [25]. Recently, however, end-to-end systems based on self-supervised learning (SSL) models have shown promising results in modelling dementia-related information in spoken language across various datasets collected from the picture description task [16], [26], [27].

The two main types of features utilised are acoustic and linguistic features. For extracting linguistic information, Bidirectional Encoder Representations from Transformers (BERT) [9], a multi-layer bidirectional transformer encoder, has

Yilin Pan is with the College of Artificial Intelligence, Dalian Maritime University, China, and with the Department of Computer Science, University of Sheffield, United Kingdom.

Bahman Mirheidari is with the Department of Computer Science, University of Sheffield, United Kingdom.

Heidi Christensen is with the Department of Computer Science, University of Sheffield, United Kingdom; and Centre for Assistive Technology and Connected Health (CATCH), University of Sheffield, Sheffield, United Kingdom, (e-mail: heidi.christensen@sheffield.ac.uk).

Daniel Blackburn is with the Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, United Kingdom.

achieved excellent performance in multiple natural language processing tasks, including linguistic-based dementia detection [23], [26], [28], [29]. For the acoustic component, wav2vec2.0 (w2v), an SSL end-to-end automatic speech recognition (ASR) system, has been employed to extract embedded acoustic information for various classification tasks, such as speaker verification [30], speech emotion recognition [31], and dementia detection [23], [26], [27]. We further explore BERT and w2v, both of which have previously demonstrated state-of-the-art performance in dementia detection [26], [29], [32], to evaluate their effectiveness in extracting acoustic and linguistic information before designing an acoustic-linguistic feature fusion system.

When constructing an automatic linguistic-based dementia detection system, an ASR system is essential for transcribing audio recordings into transcripts [16], [33]. Conventional modular ASR systems have been utilised for audio transcription in dementia detection research, yielding promising results. More recent studies have employed the pre-trained w2v system [10], which has also shown promise in generating automatic transcripts [26], [27], [34], [35]. Despite automatic transcripts often containing errors that may lead to ambiguity, previous research indicates that linguistic-based systems generally outperform acoustic-based systems [24], [33], [36]–[38], a trend attributed to the greater informativeness of linguistic features [39]. This paper first evaluates both the classic ASR system and w2v ASR system for audio transcription, then revisits the acoustic and linguistic features within the framework of state-of-the-art SSL models, demonstrating that conclusions may shift when assessed on larger datasets.

The contributions of this paper are summarised as follows: Firstly, we explore and analyse acoustic and linguistic feature extraction pipelines using w2v and BERT-based model [40], including the proposed TSAC framework to create high-performing acoustic features from w2v’s contextual layers. Secondly, we demonstrate that these features, when fused using cross-attention (TSAC-ATT), outperform various feature combination approaches. Thirdly, all experiments are conducted on two publicly available datasets (DementiaBank and ADReSS [36]) as well as our own in-house dataset (IVA), demonstrating the superior performance of the TSAC-ATT system across all datasets. By designing these experiments, we revisit inconsistent conclusions from previous research to establish a more consistent understanding of spoken language dementia detection system construction. The impact of upstream performance from ASR systems is further explored, specifically comparing the classic ASR system with the w2v ASR system, and how these affect the downstream performance of dementia classifiers. The code to replicate the results of this paper will soon be available at <https://github.com/YilinSpeechandNLP/TSAC-ATTention>.

## II. RELATED WORK

The current mainstream spoken language-based dementia detection features can be categorised into acoustic-based, linguistic-based, and acoustic-linguistic fused approaches. The remainder of this section introduces the typical linguistic and

acoustic features used for dementia detection in the literature, followed by the acoustic-linguistic feature fusion strategies.

### A. Linguistic-based Features

As dementia progresses, almost all aspects of language can be affected [3]. To represent the changes caused by dementia, researchers have proposed a bank of linguistic features to capture this, like Part-of-Speech-based [41], Type-Token-Ratio, hesitation-related features, vocabulary variation and syntactic complexity evaluation features. In recent years, deep neural networks have started to be used for extracting linguistic information directly from the transcripts. In 2019, we proposed a hierarchical attention-based system for extracting both word-level and sentence-level information for dementia detection and achieved, at the time, state-of-the-art results (74.37% F-score on automatic transcripts) on the DementiaBank (DB) dataset [16]. In the next couple of years, BERT was shown to provide superior performance for the dementia detection task.

The ADReSS [42] and ADReSSo [43] challenges, organised as part of Interspeech, aimed to provide researchers with a benchmark dataset for linguistic- and acoustic-based dementia detection tasks. In the thirteen papers published in the Interspeech-2020 ADReSS special session [36], seven papers used BERT for modelling the linguistic information [28], [29], [44]–[47]. Similarly, among all the accepted papers in the Interspeech-2021 ADReSSo special session, eight out of eleven papers employing linguistic features used BERT. Specifically, in [26], we used the BERT-based model for extracting linguistic information from the ASR hypotheses and confidence scores, which showed superior performance. In our follow-up research [48], the BERT-based model was successfully ported to the dementia regression task. Considering the superior performance of BERT in previous research, here, BERT-based system is also adopted for modelling the linguistic information embedded in the transcripts in this paper.

### B. Acoustic-based Features

A person’s voice (and therefore the acoustic aspects of their speech) is also affected by dementia, and the changes can often be seen many years before diagnosis [49], [50]. The most popular speech-based features include the Mel Frequency Cepstral Coefficient (MFCC) [51], fundamental frequency ( $F_0$ ) [52], the articulation rate, speech rate, disfluency, pause and speech rhythm features [53]–[56]. In the ICASSP 2023 Signal Processing Grand Challenge [57], a novel complementary and simultaneous ensemble algorithm on acoustic and disfluency features is proposed for acoustic feature extraction [58]. Usually, the extracted features have a relatively high dimension to ensure that symptoms are described comprehensively, such as the eGeMAPs feature set [59]. In addition to these hand-crafted features, SSL deep neural networks, like w2v, started to be used for extracting acoustic features and have shown promising results.

In the eleven papers published in the Interspeech-2021 ADReSSo special session, the pre-trained w2v system [10] was used by four papers [26], [27], [34], [35]. The w2v model

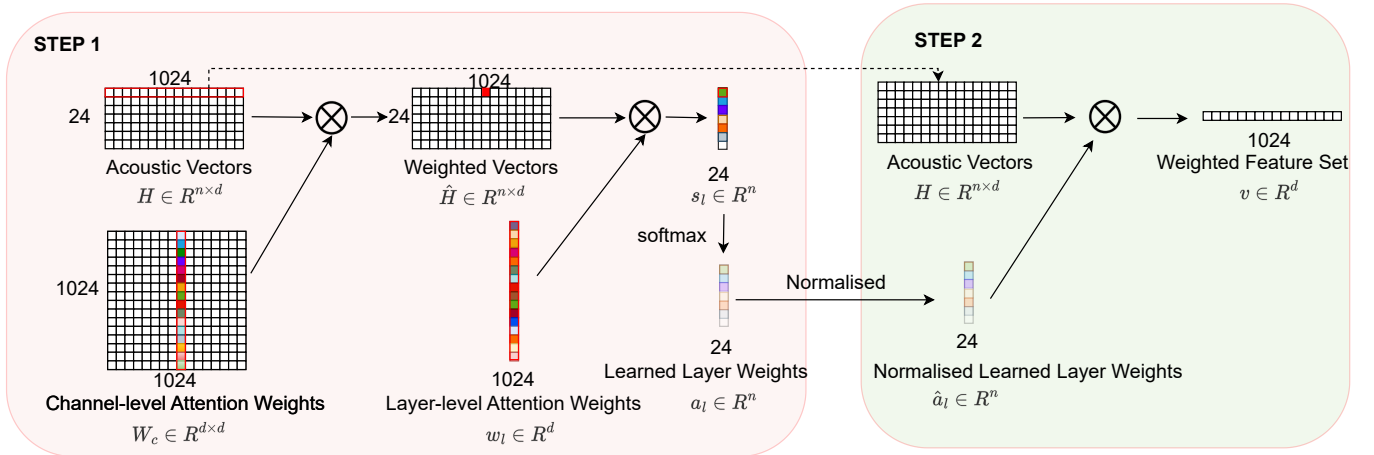


Fig. 1: Diagram of the proposed two-step attention-based feature combination (TSAC) framework for combining the multiple acoustic vectors into one single acoustic feature set.

is comprised of three parts: a Convolutional Neural Network-based (CNN-based) local encoder, multiple contextualised representations with transformers (contextual layers), and a quantisation module. The vectors output by the contextual layers have been used as the acoustic representation. A layer-wise method was proposed for selecting the vector output, among the multiple contextual layers, as the layer that performs the best on a development set [26]. Leonardo et al [31] proposed to combine the output of different w2v's contextual layers by using trainable weights learned jointly with the classification task. In this paper, both of these two methods are replicated to explore how to select the acoustic features generated by w2v.

### C. Acoustic-linguistic Feature Fusion

In evaluating both acoustic and linguistic features, it is crucial to consider the method of fusion, as previous research has yielded inconsistent conclusions. Some studies suggest that integrating acoustic and linguistic information may not be necessary [32], [60], while others advocate for joint modelling of these modalities [36], [37], [47], [61]–[64]. This paper will explore whether integrating acoustic and linguistic information is necessary and how the two types of features should be integrated. Early fusion, which involves concatenating features from both modalities before classification, is a prevalent approach [39], [65]. However, its impact on classification performance is variable and does not always surpass the use of single modalities [26], [47]. To address this, cross-attention has been proposed to balance modalities by using an attention mask to emphasise features from one modality in another before fusion [66]–[68]. This paper investigates the application of cross-attention in the TSAC-ATT system for dementia detection, focusing on the fusion of acoustic and linguistic information.

## III. TWO-STEP ATTENTION-BASED FEATURE COMBINATION CROSS-ATTENTION SYSTEM

This section presents the SSL-based TSAC-ATT system, which integrates vectors extracted from the contextual layers of w2v into a unified acoustic vector. This vector is

subsequently fused with the linguistic feature vector derived from BERT outputs. The system is comprised of two components: the TSAC feature combination framework and the cross-attention feature fusion framework. Figure 1 illustrates the TSAC feature combination framework, while Figure 2 depicts the complete TSAC-ATT system. The design of the TSAC framework aims to enable the flexible integration of multiple acoustic vectors from the w2v contextual layers into a consolidated acoustic feature set, which incorporates learned attention weights for subsequent feature fusion.

### A. TSAC Feature Combination Framework

As illustrated in Figure 1, the proposed TSAC feature combination framework consists of two steps. First of all, the outputs from the multiple contextual layers of w2v form an acoustic feature tensor of size  $[n, T, d]$ , where  $T$  varies with the length of the input audio recordings,  $n$  represents the number of contextual layers in w2v, and  $d$  denotes the feature dimension. For our system,  $n$  equals to 24 and  $d$  equals to 1024. To produce fixed-length acoustic vectors  $H \in [24, 1024]$ , the acoustic feature tensor is averaged over time, following the method detailed in [26]. As shown in the figure, in STEP 1, a channel-level attention mechanism is then applied to acoustic vectors  $H$  to calculate channel weights for each dimension, yielding channel-weighted acoustic vectors  $\hat{H} \in [24, 1024]$ . For example, as illustrated in the figure, the ninth dimension of the channel-weighted vector from the first contextual layer (highlighted in red) is derived by multiplying the vector from the first contextual layer by the corresponding ninth dimension vector in the channel-level attention weights matrix. Then, layer-level attention weights are applied to the channel-weighted vectors to compute the learned layer weights  $\alpha_l$ . The equations describing STEP 1, as illustrated in Figure 1, are as follows:

$$\begin{aligned} \hat{h}_i &= \tanh(W_c h_i + b_c) \\ s_i &= \tanh(w_l \hat{h}_i + b_l) \\ \alpha_i &= \text{softmax}(s_i) \end{aligned} \quad (1)$$

where  $\mathbf{h}_i$  represents the  $i_{th}$  vector extracted from the  $i_{th}$  contextual layer of w2v’s acoustic vectors  $\mathbf{H}$ . The channel-level attention weight matrix and bias are denoted by  $\mathbf{W}_c$  and  $b_c$ , respectively, while  $\mathbf{w}_l$  and  $b_l$  refer to the layer-level attention weight vector and bias. These parameters are initialised randomly.  $\tanh$  is used as the non-linear function. The term  $\alpha_i \in \mathbf{a}_l$  represents the learned layer weight vector before normalisation.

After obtaining the normalized learned layer weights  $\hat{\alpha}_i \in \hat{\mathbf{a}}_l$ , STEP 2 combines the acoustic vectors  $\mathbf{H}$  with the normalized learned layer weights  $\hat{\mathbf{a}}_l$  to produce the layer-weighted feature set  $\mathbf{v}$ :

$$\mathbf{v} = \mathbf{a}_l^\top \mathbf{H} \quad (2)$$

The proposed TSAC framework learns the 24-dimensional layer weights by incorporating both channel-level and layer-level information from the acoustic vector sets  $\mathbf{H}$  for each audio recording, which is designed to be trained in conjunction with the cross-attention system described in Section III-B.

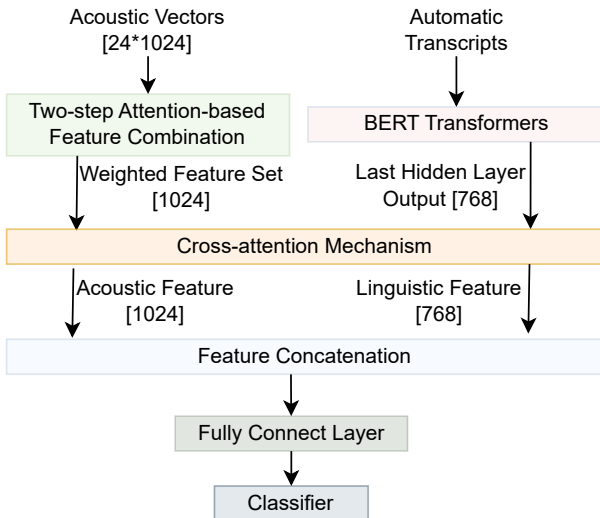


Fig. 2: The diagram of the proposed TSAC cross-attention-based feature fusion system (TSAC-ATT).

### B. TSAC-ATT Feature Fusion System

The complete TSAC-ATT system is shown in Figure 2, while the green block corresponds to the structure shown in Figure 1. The acoustic vectors  $\mathbf{H}$  are processed by the TSAC framework shown in Figure 1 for getting the layer-weighted acoustic feature set  $\mathbf{v}$ , which is used for combining with the linguistic features extracted from BERT. Similar to [26], [48], the output feature matrix of BERT’s last hidden layer is averaged across time and used as the linguistic feature set  $\mathbf{u}$ . To fuse the layer-weighted feature set  $\mathbf{v}$  with the linguistic feature set  $\mathbf{u}$ , the cross-attention technique is used. Specifically, to complement the information embedded in different modalities, cross-attention is applied by creating query (q)  $\mathbf{x}_q$  from modality A (namely the acoustic feature

set), and key (k)  $\mathbf{x}_k$ , value (v)  $\mathbf{x}_v$  from modality B (namely the linguistic feature set), as shown in Equation 3:

$$\mathbf{z} = \text{softmax} \left( \frac{f_q(\mathbf{x}_q) f_k(\mathbf{x}_k)}{\sqrt{d_k}} \right) f_v(\mathbf{x}_v) \quad (3)$$

where  $f_q$ ,  $f_k$ , and  $f_v$  denote the linear layers for the query  $\mathbf{x}_q$ , key  $\mathbf{x}_k$ , and value  $\mathbf{x}_v$  components, respectively. The feature dimension is represented by  $d_k$ . The variable  $\mathbf{z}$  refers to the processed feature, which can be either the acoustic or linguistic feature set as illustrated in Figure 2. The cross-attention processed feature sets from the two modalities are concatenated. All the parameters in the TSAC-ATT system, as shown in Figure 2, are trained jointly.

## IV. DATASET INFORMATION

The picture description task is a widely used method for dementia detection, focusing on semantic knowledge and retrieval memory [13]. The ‘‘Cookie Theft’’ line drawing, originally created for aphasia testing [69], is frequently used as a prompt. Participants are asked to describe the picture, and their responses are recorded for subsequent evaluation by a neuropsychologist. This paper uses three datasets for the experimental work, including publicly available datasets (DB and ADReSS) and a dataset named IVA, (the abbreviation of Intelligent Virtual Agent) collected by our collaborators at the Royal Hallamshire Hospital (Sheffield, United Kingdom). These datasets include audio recordings of picture descriptions using the Cookie Theft picture, corresponding manual transcripts, and clinically obtained diagnostic labels.

### A. DementiaBank Dataset

With a total of 551 samples, it stands as the largest publicly available speech dataset for evaluating cognitive impairment. As detailed in Table I, the dataset comprises 222 samples from 89 healthy controls (HCs) and 255 samples from 168 individuals with Alzheimer’s Disease (AD). The remaining samples are from individuals with other forms of dementia or those who transitioned from Mild Cognitive Impairment (MCI) to AD during the data collection period. For the purposes of this paper, only recordings from the AD and HC groups are utilized, totalling 477 recordings.

TABLE I: Gender, age and duration (given in seconds) statistics for the DB dataset.

Patient group	Gender (M:F)	Average Age	Average Duration
AD	85:170	71.60±(8.41)	56.14±(23.77)
HC	79:143	64.17±(7.99)	59.28±(31.40)
Others	44:30	68.29±(9.31)	54.82±(22.97)

### B. ADReSS Dataset

The ADReSS dataset, created for the Interspeech-2020 challenge, was derived from a subset of the DB dataset to achieve better balance in diagnostic classes, age, and gender. To address the challenging audio quality of the DB

dataset—characterised by high levels of environmental noise and variable microphone placements—the ADReSS organisers improved the recordings through noise removal and volume normalisation. The information about the ADReSS dataset [42] is shown in Table II. As opposed to the DB and IVA datasets, the ADReSS dataset has separate test sets. In the training set, there are 54 recordings from the HC and AD groups, respectively. The test set also includes a balanced number of recordings from the HC and AD groups; 48 recordings in total.

TABLE II: Gender, age and duration (given in seconds) statistics for the ADReSS dataset.

Subset	Patient group	Gender (M:F)	Average Age	Average Duration
Training	AD	24:30	66.91±(6.52)	82.24±(43.21)
	HC	24:30	66.21±(6.41)	61.46±(20.76)
Test	AD	11:13	66.13±(7.28)	90.47±(51.75)
	HC	11:13	66.13±(6.94)	74.55±(31.51)

### C. IVA Dataset

The IVA dataset comprises recordings of individuals responding to memory-probing questions and performing verbal fluency tests designed to simulate those used by neurologists in clinical settings. For consistency with the other datasets, this paper uses only the "Cookie Theft" picture description portion of the IVA dataset. As detailed in Table III, the dataset includes a total of 62 recordings from healthy controls (HCs) and 29 recordings from individuals with Neurodegenerative Disorders (NDs), which encompass Alzheimer's Disease (AD), Vascular Dementia, and Parkinson's Disease. The audio quality of the IVA recordings is superior to that of the DB dataset [70], primarily due to the recordings being more recent (from 2016 onwards) [71].

TABLE III: Gender, age and duration (given in seconds) statistics for the IVA dataset; UNK is used to represent the unknown gender.

Patient group	Gender (M:F:UNK)	Average Age	Average Duration
HC	6:10:46	70.56±(8.44)	75.12±(33.61)
ND	12:10:7	69.77±(6.62)	69.38±(40.57)

### D. Data Segmentation

As shown in Table I, Table II and Table III, which is too lengthy for direct use in fine-tuning pre-trained w2v models. Consequently, the recordings are segmented into segments corresponding to the length of each sentence. The total number of sentences and their average durations are detailed in Table IV for each dataset. For the DB dataset, segmentation is performed using the provided start and end times for each sentence. The sentence-level audio recordings provided by the ADReSS dataset, which have been enhanced for speech quality, are used directly. In contrast, the IVA dataset is segmented manually.

TABLE IV: Number of sentences and the average durations for the three processed datasets (duration given in seconds).

Dataset	# Sentences	Average Duration
DB	5972	4.53
IVA	1208	6.53
ADReSS	4077	4.59

## V. EXPERIMENTAL SETUP

This section first introduces the cross-validation (CV) methodology employed to evaluate the proposed TSAC-ATT system. Next, the ASR systems utilised for audio transcription are detailed. Following this, the parameters for the linguistic feature extraction systems, acoustic feature extraction systems, and fusion systems are outlined.

### A. Cross Validation Setting

CV was employed to train the dementia detection systems, given the relatively small size of the datasets. For the DB and IVA datasets, a "speaker-independent" 10-fold CV approach was used to ensure that no speaker appeared in both the training (8 folds), testing (1 fold), or development (1 fold) sets simultaneously. The CV lists for the DB dataset are available on GitHub<sup>1</sup> and align with those used in [16], [24]. The corresponding results reported in Section VI are the averaged result of the 10-fold test sets. For the ADReSS dataset, the 108 speakers in the training set were divided into 9 folds of 12 speakers each, as in [37]. Since the test set is fixed, the reported test set results in Section VI are based on majority voting from the predictions across the 9 folds.

### B. ASR Systems Setting

To transcribe the audio recordings into text, two ASR systems were evaluated: a classic system and an end-to-end system. The classic ASR system utilised a Kaldi Librispeech recipe [72], which provided a pre-trained time delay neural network acoustic model. This model was then fine-tuned using the transfer learning technique described by [73] (transferring all layers). The 10-fold CV methodology as described in Section V-A was employed to fine-tune the ASR systems across all corresponding datasets (described in Section IV), to generate the automatic transcripts. The end-to-end ASR system utilised was the large model of w2v, which comprises 24 transformer blocks with 16 attention heads, initialised with the pre-trained model *Facebook/wav2vec2-large-960h-lv60-self*. A 10-fold CV was used for transcribing the audio recordings into texts. For fine-tuning, the following parameters were used: 20 epochs, a batch size of 1, and a learning rate of 1e-5. The audio files were segmented into sentences as described in Section IV, and the transcribed sentences were concatenated with period punctuation ("."). The trained model was selected based on the Word Error Rate (WER) obtained on the development set across the 10 folds.

Table V shows the WER of the two ASR systems. To assess the impact of acoustic noise reduction applied to the

<sup>1</sup><https://github.com/YilinSpeechandNLP>

TABLE V: The WER of the two ASR systems on the different datasets.

Test set	Classic	w2v
DB	33.19	48.43
IVA	25.31	35.06
ADReSS	52.33	57.31
DB-subset	37.42	50.66

DB subset used in ADReSS, we separately evaluated the ASR performance on this subset (DB-subset). The results show that the WER of transcripts generated by w2v is consistently higher than that from the conventional classic ASR system across all datasets. **Since we used w2v without feeding any language models data to the network, the word error rate is still high. As a further work we could try to use both acoustic and linguistic information from our training dataset.** Notably, the noise-reduced files in the ADReSS dataset are the most challenging to recognise, while the corresponding DB-subset files achieve WERs more comparable to those of the original DB dataset. As noted, the IVA dataset, with its superior audio quality, results in the best WER. **In this paper, the audio recordings from the DB, ADReSS and IVA datasets are not further denoised. In our current work [74], we are working on designing an AD specific speech enhancement system for improving the denoised audio recording's performance in the downstream dementia detection system.**

### C. Linguistic-only System

BERT-for-Sequence-Classification is composed of multiple transformer layers and a two-dimension fully connected layer (here referred to as BERT). It has successfully been used for modelling the linguistic information and doing classification in previous research [26], [29]. In this paper, it is trained to evaluate the manual transcripts and automatic transcripts generated by the ASR systems described in Section V-B. The BERT base model that includes twelve layers of transformers block with a hidden size of 768 and a number of self-attention heads as twelve was initialised by <https://github.com/huggingface/transformers>. **The parameters in the BERT pre-trained model are fine-tuned with the transcripts generated by ASR on the AD classification task with the CV settings described in Section V-A.** The parameters were set as below: the number of epochs was set to 8, the batch size was set to 4, and the max transcript length was set to 256. Transcripts with a longer length were chunked, and shorter transcripts were padded. These parameters were set according to the performance of different datasets on the development set.

### D. Feature Fusion System Setting

In this paper, the dimension of feature vector equals to 1024, which corresponds to the dimension of the w2v contextual layers. After processing by the designed TSAC framework, the output from 24 contextual layers results in 24 vectors. These extracted acoustic vectors are utilised as the input for the TSAC framework. The BERT Transformers, featuring twelve transformer layers, a hidden size of 768, and twelve

self-attention heads, was initialised using <https://github.com/huggingface/transformers>. For the cross-attention mechanism (ATT) introduced in Section III-B, the number of multi-head attention nodes was set to 8, and the head dimension was set to 64. As shown in Figure 1, for the TSAC framework, the channel-level attention matrix  $W_c \in [1024, 1024]$  and layer-level attention vector  $w_l \in [1, 1024]$  were initialised randomly. The fully connected layer was used to reduce the 1792-dimension concatenated feature into 256 dimensions before doing classification. To train the feature fusion system, the batch size was set to 4, and the number of epochs was set to 8. The maximum length of each transcript was set to 256. These parameters were set according to the performance of the development set.

### E. Baseline Systems Setting

The TSAC framework is designed to combine multiple acoustic vectors into a feature set, to be used in the subsequent feature fusion system. In order to evaluate the proposed TSAC feature combination framework, five baseline feature processing systems were implemented as well. They can be categorised into feature combination or feature selection approaches. These are described below:

- **Random:** a feature selection approach designed to select one acoustic vector out of multiple acoustic vectors.
- **Last-layer:** a feature selection approach which uses the acoustic feature output by the w2v's last hidden layer as the acoustic feature used for feature fusion [34], [35].
- **Acoustic-select:** also a feature selection approach. It uses the acoustic feature output by the w2v's that achieved the best result in the acoustic-only system as the acoustic feature in the feature fusion system.
- **Layer-wise:** The layer-wise feature fusion method selects the vector output by contextual layers that has shown the best performance on a development set [26].
- **Weighted:** A feature combination method. The 24 acoustic vectors are combined using Equation 2, where the  $\alpha_i$  variables are randomly initialised.

## VI. RESULTS

**In this section, we present the results obtained from evaluating our proposed systems. Firstly, in Section VI-A, We compare the performance of the linguistic-only model against both manually created transcripts and transcripts generated the two ASR systems. Then, in Section VI-B, we assess the efficiency of both the proposed TSAC framework and the complete TSAC-ATT system using the datasets outlined in Section IV.**

### A. Results with Different Transcripts

**To compare the manual transcripts and the transcripts generated by the commonly used ASR systems, namely the end-to-end and classic ASR systems, the BERT base model described in Section V-C is fine-tuned and utilised as the linguistic-only system.** By comparing the transcripts generated by the two ASR systems (the first and second lines), it is found

TABLE VI: The linguistic-only results (%) by using the fine-tuned BERT base system on the automatic transcripts generated by the two ASR systems and manual transcripts.

Dataset	Transcript	Accuracy	Precision	Recall	F-score
DB	Classic ASR	80.43	80.70	81.26	79.74
	w2v ASR	80.04	80.34	80.04	80.06
	Manual	<b>81.76</b>	<b>82.64</b>	<b>81.76</b>	<b>81.76</b>
IVA	Classic ASR	85.39	85.84	85.39	84.43
	w2v ASR	<b>86.36</b>	<b>86.09</b>	<b>86.36</b>	<b>86.01</b>
	Manual	84.27	83.96	84.27	83.65
ADReSS	Classic ASR	70.14	71.04	70.14	69.84
	w2v ASR	<b>77.08</b>	77.13	<b>77.08</b>	<b>77.07</b>
	Manual	<b>77.08</b>	<b>77.51</b>	<b>77.08</b>	76.99
DB-subset	Classic ASR	73.84	76.24	83.84	73.22
	w2v ASR	<b>82.61</b>	<b>82.61</b>	<b>82.61</b>	<b>82.61</b>
	Manual	77.08	77.51	77.08	76.99

that the transcripts from w2v, though having a significantly higher WER (by referring to the results in Table V), actually gives a marginally better performance than the transcripts from the classic ASR system. For example, for the DB dataset, the WER of the transcripts generated by the classic ASR system is 33.19%, corresponding to a 79.74% F-score. In comparison, the WER of transcripts generated by w2v is much higher 48.43%, but the corresponding F-score is 80.06%. In our current work [75], [76], it is demonstrated that some information related to AD classification is embedded in the w2v’s automatic transcripts. It is inferred that some AD-related information embedded in the w2v’s automatic transcripts are learned by our designed system, making the performance of the automatic transcripts better than the manual transcripts.

Furthermore, the manual transcripts are also used as the input of the BERT system, as shown in the last line of each dataset in Table VI. By comparing the results from automatic and manual transcripts, it is found that the automatic transcripts generated by w2v can perform better than or similar to the manual transcripts when being used as the input of BERT. Specifically, on the IVA dataset, the F-score is 86.01% with automatic transcripts from the w2v system, compared to 83.65% F-score with manual transcripts. Informed by the results shown in Table VI, the transcripts generated by the w2v are used in the following up experiments for evaluating the feature fusion system.

### B. Results on the TASC-ATT System

As discussed in Section III, TSAC-ATT system include two parts: the TSAC framework and the ATT module. In this section, the efficiency of the proposed TSAC framework is analysed first. To this end, the TSAC framework is trained by concatenating directly with the linguistic feature output by BERT for feature fusion, which is similar to TSAC-ATT but without the ATT part. The five feature combination or selection methods described in Section V-E works as the baseline systems. The corresponding results are shown in Table VII. The best result for each dataset is indicated with **bold**, and the second best result is indicated with underline.

As shown in Table VII, the proposed TSAC framework performs the best or second best on all the datasets. The

TABLE VII: Results (%) using different feature combination or selection methods for acoustic feature generation. The fused features is generated by feature concatenation.

Dataset	Method	Accuracy	Precision	Recall	F-score
DB	Random	78.57	78.60	78.57	78.58
	Last-layer	72.06	72.13	72.06	72.08
	Acoustic-select	80.76	80.75	80.82	80.81
	Layer-wise	<b>81.35</b>	<b>81.35</b>	<b>81.35</b>	<b>81.35</b>
	Weighted	77.94	77.98	77.94	77.95
	TSAC	<u>80.87</u>	<u>80.85</u>	<u>80.87</u>	<u>80.89</u>
IVA	Random	86.36	86.09	86.36	86.01
	Last-layer	86.52	86.28	86.52	86.19
	Acoustic-select	85.63	85.57	85.59	85.64
	Layer-wise	<b>89.74</b>	<b>90.21</b>	<b>89.74</b>	<b>89.28</b>
	Weighted	84.09	83.68	84.09	83.68
	TSAC	<u>88.64</u>	<u>88.76</u>	<u>88.64</u>	<u>88.16</u>
ADReSS	Random	77.08	77.13	77.08	77.07
	Last-layer	72.45	73.72	72.45	72.62
	Acoustic-select	75.41	74.98	74.40	75.43
	Layer-wise	76.39	76.50	76.39	76.36
	Weighted	75.00	75.17	75.00	74.96
	TSAC	<b>79.17</b>	<b>79.37</b>	<b>79.17</b>	<b>79.13</b>
DB-subset	Random	79.17	79.37	79.17	79.13
	Last-layer	73.15	73.47	73.15	73.19
	Acoustic-select	79.30	79.30	79.30	79.30
	Layer-wise	79.40	79.54	79.40	79.37
	Weighted	<b>83.33</b>	<b>85.57</b>	<b>83.33</b>	83.30
	TSAC	<b>83.33</b>	<u>83.33</u>	<b>83.33</b>	<b>83.33</b>

two feature selection methods (using the vector output by w2v’s last layer or selecting the layer randomly), are tested but overall showed an inferior performance to the fusion-based approaches. In comparison, the layer-wise feature fusion system got the best result on the IVA and DB datasets. However, all the systems except the layer-wise feature fusion system only need to be trained once using one acoustic vector, whereas the layer-wise feature fusion needs to be trained 24 times by individually using 24 acoustic vectors (more analyse provided in Section VII). The result shown in Table VII reveals that TSAC is a superior acoustic feature processing framework for getting a feature vector used for the feature fusion system by considering both the performance and the time cost.

The results of using the TSAC-ATT system are presented in Table VIII. Compared to the results that directly fuse acoustic and linguistic features through concatenation, as shown in Table VII, the results in Table VIII indicate that the cross-attention mechanism contributes to the superior performance of the proposed TSAC-ATT system. For example, the F-score improves from 79.13% to 81.24% on the ADReSS dataset after incorporating cross-attention into the TSAC framework. Additionally, when compared to the linguistic-only results presented in Table VI, our TSAC-ATT system demonstrates superior performance across all provided datasets.

TABLE VIII: The results (%) of the TSAC-ATT system.

Dataset	Accuracy	Precision	Recall	F-score
DB	81.51	81.66	81.51	81.53
IVA	88.64	88.76	88.64	88.16
ADReSS	81.25	81.30	81.25	81.24
DB-subset	83.33	83.33	83.33	83.33

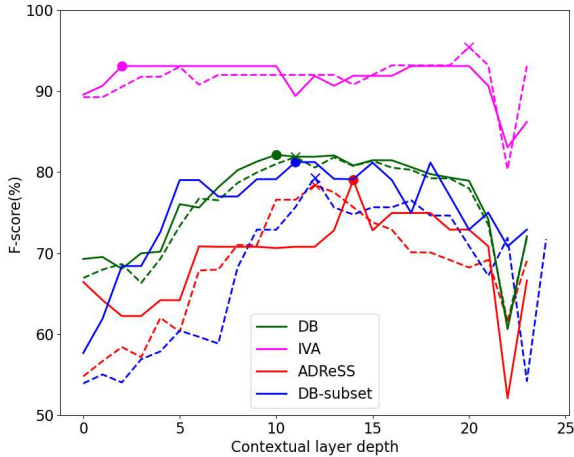


Fig. 3: The relationship between the contextual layer depth and F-scores using the vectors extracted from the corresponding layer on the development set and test set. Dash line and solid line are used to represent the result from the test set and development set; real point and cross are used to represent the best result in the test set and development set respectively.

### VII. ABLATION STUDY

In this section, we first outline the motivation behind designing the TSAC framework. Next, we provide a comprehensive comparison of related systems using various modality inputs to illustrate the robustness of our proposed TSAC-ATT system. Finally, we visualise the layer weights obtained through two feature combination methods, specifically, the layer-wise and TSAC, to enhance the interpretability of the learned parameters.

#### A. Motivation for Designing the TSAC framework

The motivation for designing the TSAC framework arises from the varying performance of acoustic features extracted from the different contextual layers of w2v. To explore the relationship between the vectors extracted from these layers and their performance in AD detection, this section uses the output vectors from w2v’s contextual layers as acoustic features in a pipeline system, classified by a TB classifier, as described in [26] (here referred to as the acoustic-only system). The F-score results for each layer across the development and test sets are shown in Figure 3.

In previous research, the vector extracted from the last contextual layer has been used as input for back-end classifiers [26], [34], [35]. However, as illustrated in the figure, relying solely on the last contextual layer does not guarantee superior performance in the acoustic-only dementia detection system. Analysis of different datasets reveals a similar trend among the DementiaBank-related datasets (DB, ADRess, and DB-subset), where the middle layers demonstrate higher performance. Also, as indicated by the results in Table VII, selecting a specific layer as the acoustic feature vector cannot ensure optimal performance across datasets for the fusion system.

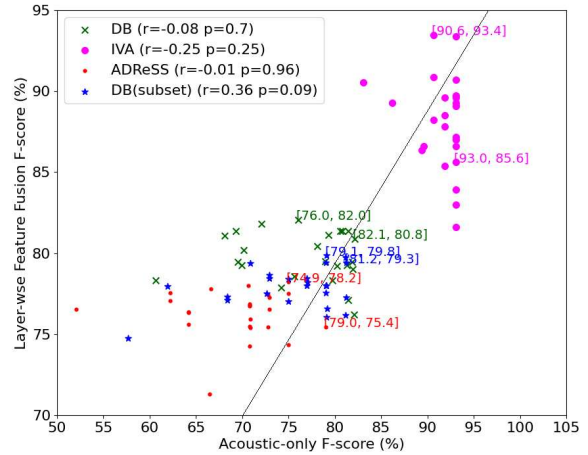


Fig. 4: The relationship of F-scores (%) achieved by the acoustic-only and layer-wise feature fusion systems using the vectors extracted from the w2v’s contextual layers on the four datasets’ test set.

While the layer-wise feature selection method ensures comparable performance on the feature fusion system (as shown in Table VII), it is time-consuming. One idea is selecting the acoustic vector extracted from a specific layer based on its acoustic-only result, which is more straightforward. To this end, the relationship between the F-scores of the acoustic-only and layer-wise feature fusion systems using the vector extracted from the same contextual layer is shown Figure 4. Only the locations of the best acoustic-only and feature fusion results are marked. The figure also presents the Pearson correlation ( $r$ ) and  $p$ -value between the acoustic-only and feature fusion systems. With a significance level set at a  $p$ -value of 0.05, no significant correlation exists between the performance of the acoustic-only system and that of the feature fusion system using the same acoustic vector. In other words, selecting an acoustic vector for layer-wise feature fusion remains challenging depending on the performance of the acoustic-only system.

To sum up, the vectors extracted from the contextual layers of w2v can deliver superior performance for dementia detection. However, the relationship between the performance of multiple acoustic vectors from different layers and their depth varies across datasets. This variability motivates us to design a system that incorporates and processes multiple vectors as a consistent representation.

#### B. Results Comparison

**Acoustic-only Comparison:** The published acoustic-only results of the datasets used in this paper are summarised in Table IX, together with our results. Our results are the test set results corresponding to the contextual layer that exhibits the best performance on the development set (shown in Figure 3). As shown, the performance of our acoustic-only system on all the datasets is excellent. Specifically, the F-score is 82.54% using the acoustic feature only, which is superior to previous

research on the DB dataset that uses acoustic information only. In comparison, an accuracy of 78.70% [77] and 68.60% [38] was respectively achieved by using popular acoustic feature sets. In [76], an accuracy of 79.04% was achieved by learning the disfluency information represented by the path signature of the acoustic features.

Compared to the previous research on the ADRess dataset, the best acoustic-only F-score reported by the papers accepted by Interspeech-2020 ADRess special session was 72.62% using the pre-trained VGGish system [78] for acoustic information extraction [47], compared to 70.78% F-score and 81.24% F-score achieved in Table IX on the ADRess and DB-subset datasets, respectively. Compared to our acoustic-only method, which utilises a simple linear classifier for classifying the w2v features, the method proposed by [47] used not only the pre-trained VGGish model for feature extraction, but also more complex classifier (a modified version of Convolutional Recurrent Neural Network) for feature classification.

TABLE IX: The acoustic-only results (%) in previous research. PRE: Precision, REC: Recall, FS: F-score, ACC: Accuracy.

Dataset	Method	ACC	PRE	REC	FS
DB	Mittal et al. [38]	68.60	-	-	-
DB	Haider et al. [77]	78.70	-	-	-
DB	Pan et al. [76]	79.04	-	-	78.91
DB	Ours	<b>82.56</b>	<b>82.56</b>	<b>82.56</b>	<b>82.54</b>
ADReSS	Koo et al. [47]	72.92	73.96	72.92	72.62
ADReSS	Ilias et al. [79]	63.33	66.01	55.83	60.30
ADReSS	Ours	70.83	70.98	70.83	70.78
DB-subset	Cummins et al. [37]	63.90	-	-	63.90
DB-subset	Ours	<b>81.25</b>	<b>81.30</b>	<b>81.25</b>	<b>81.24</b>

**Linguistic-only Comparison:** The published linguistic-only results of the datasets used in this paper are summarised in Table X, together with our linguistic-only results, using the w2v as the automatic transcripts, and BERT as the classifier. Though the hierarchical attention approach proposed in [16] achieved state-of-the-art results on the DB dataset when published, this is superseded by BERT on all the datasets. Compared to previous research, The research reported in [37] achieved an F-score of 81.30% with automatic transcripts, whereas our linguistic-only system obtained 82.61% F-score on the DB-subset and 77.07% F-score on the ADRess dataset. Additionally, the result presented in [38] showed 74.50% F-score using automatic transcripts on the DB dataset, compared to 80.06% F-score in this study.

**Modality Comparison:** The acoustic-only, linguistic-only and feature fusion results of this paper on all the datasets are shown in Figure 5. As shown, under the same condition, the performance of different systems on the DB-subset dataset is always better than the ADRess dataset, though the noise level of the audio recordings in the DB-subset is higher than in the ADRess dataset. These results demonstrate that the acoustic enhancement method used in [36] has a detrimental effect on the both ASR and classification performance.

As illustrated in Figure 5, the best result on the DB dataset is 82.54% F-score with the acoustic-only system, compared to 80.06% for the linguistic-only system and 81.53% for the TSAC-ATT feature fusion system. Similarly, the best result

TABLE X: The linguistic-only results (%) by using the hierarchical attention-based system and BERT on the automatic transcripts generated by the classic ASR system. PRE: Precision, REC: Recall, FS: F-score, ACC: Accuracy.

Dataset	System	ACC	PRE	REC	FS
DB	Pan et al. [16]	76.76	76.73	76.76	76.74
DB	Mittal et al. [38]	75.50	-	-	74.50
DB	Ours	<b>80.04</b>	<b>80.34</b>	<b>80.04</b>	<b>80.06</b>
ADReSS	Pan et al. [16]	68.75	69.16	69.90	68.61
ADReSS	Ours	77.08	77.13	77.08	77.07
DB-subset	Cummins et al. [37]	81.30	-	-	81.20
DB-subset	Ours	<b>82.61</b>	<b>82.61</b>	<b>82.61</b>	<b>82.61</b>

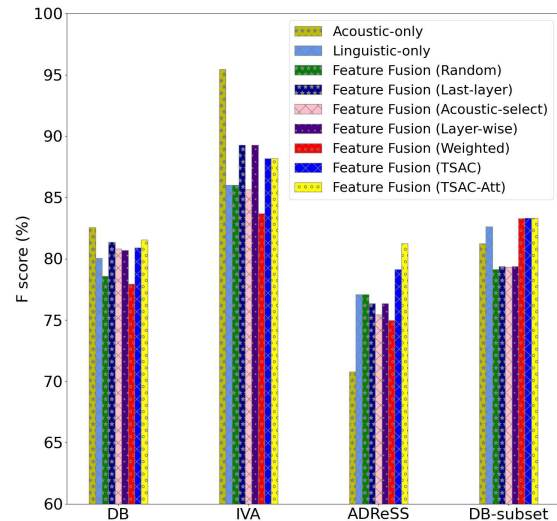


Fig. 5: The results (F-score) from the acoustic-only, linguistic-only and feature fusion by using the acoustic features and/or transcripts output by w2v.

on the IVA dataset is 95.45% F-score from the acoustic-only system, while the linguistic-only system achieves 84.43% and the layer-wise feature fusion system reaches 89.28%. The noticeable gap between the acoustic-only and feature fusion systems is partly due to the high quality of audio recordings in the IVA dataset, which ensures the extraction of high-quality acoustic features from w2v. Previous studies have shown that feature fusion systems often struggle to match the performance of linguistic-only systems [32], [47], [60], but this is the first study where the acoustic-only system shows similar limitations. Encouraged by these findings, future research should investigate how high-quality acoustic information embedded in audio recordings can be extracted directly, without relying on ASR-generated automatic transcripts.

### C. Learned Layer Weights Analysis

To understand the two acoustic vector combination methods, specifically the weighted feature combination and TSAC feature combination, the learned layer weights extracted from

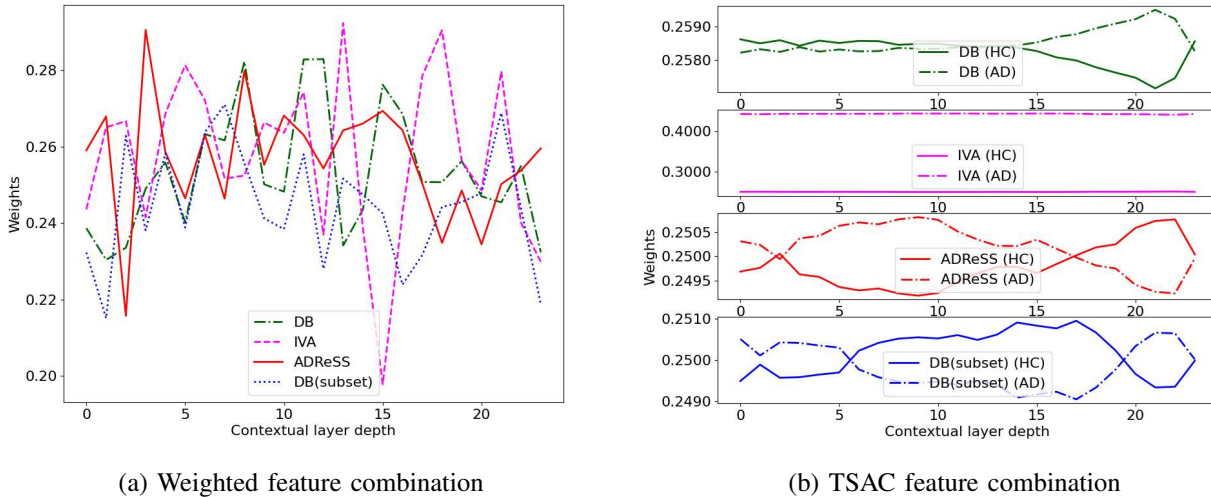


Fig. 6: The learned weights of the weighted feature fusion method proposed in [31] (left) and the learned weights from HC and AD/ND with the proposed TSAC framework (right).

the weighted feature fusion [31] and the TSAC framework are shown in Figure 6. The learned weights using the weighted feature fusion method are shown in Figure 6 (a), which is difficult to interpret. In comparison, the learned layer weights extracted from the trained TSAC framework are plotted in Figure 6 (b). The weights learned for each acoustic feature set correspond to each audio recording, meaning the weights are distinctive for each recording. The recordings' learned weights are averaged for the two classes (HC and AD/ND). As shown, the layer-level attention mechanism can learn the acoustic difference between the recordings collected from the HC and AD/ND.

## VIII. CONCLUSION AND FUTURE WORK

This paper explored the application of robust SSL models, specifically BERT and w2v, in the context of dementia detection using two publicly available datasets as well as our own in-house dataset, all centered on Healthy Controls and individuals with dementia describing the Cookie Theft picture. A secondary aim of the paper was to revisit the common conclusion drawn in similar studies that acoustic-only systems generally perform inferiorly compared to linguistic-based or multimodal (acoustic and linguistic) systems.

The analysis of the correlation between the 24 contextual layers of the w2v model and dementia detection performance revealed no clear relationship between the performance of vectors from different layers and their effectiveness in the layer-wise feature fusion system. Notably, the best-performing vector in the acoustic-only system did not yield optimal results in the layer-wise feature fusion system. This finding motivated the development of the Two-Step Attention-based Feature Combination (TSAC) framework and its integration with a cross-attention-based feature fusion (TSAC-ATT) system. Comparisons among TSAC, TSAC-ATT, and various baseline feature fusion systems demonstrated that both the TSAC

framework and the cross-attention mechanism significantly enhance the performance of the proposed TSAC-ATT system.

A key finding highlights the impact of ASR system performance on downstream classifiers. Our exploration of two ASR systems, namely the classic and the end-to-end w2v, revealed that despite the classic system generating transcripts with a lower WER across all datasets, the w2v system outperformed it in dementia detection. This aligns with previous work [26], which demonstrated that using multiple ASR hypotheses and their confidence scores could yield superior results, even when WER was not minimised. Thus, relying solely on WER to evaluate ASR systems is inadequate for optimising downstream performance. The acoustic-only SSL-based system achieved outstanding results across three datasets, suggesting the potential for developing high-performance, non-ASR dependent systems. Such advancements could enable the creation of more dialect- and accent-agnostic systems, crucial for fair application in clinical settings without bias toward language background [80].

Future work will delve into why w2v-generated transcripts, despite a relatively high WER, can outperform other systems in downstream dementia detection. Additionally, we aim to develop more straightforward SSL systems for directly extracting acoustic and linguistic information from audio recordings. **Finally, although this work presents a very comprehensive evaluation across all publicly available dementia datasets, exploring how these may be combined would also be of interest (cross-corpora, cross-language and cross-disease experimentation).**

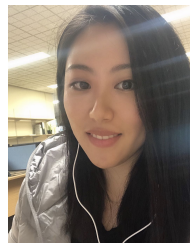
## ACKNOWLEDGMENTS

This work is supported under the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Training Network for PAthological Speech processing; Grant Agreement No. 766287).

## REFERENCES

- [1] Alzheimer's Disease International, "Dementia statistics," 2022, <https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/> [Online].
- [2] H. E. Hamilton, *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge University Press, 2005.
- [3] K. Groves-Wright, J. Neils-Strunjas, R. Burnett, and M. J. O'Neill, "A comparison of verbal and written language in Alzheimer's disease," *Journal of Communication Disorders*, vol. 37, no. 2, pp. 109–130, 2004.
- [4] J. Sundberg and R. Sataloff, *Vocal tract resonance*. Plural Publishing San Diego, California, 2005.
- [5] R. L. Horwitz-Martín, T. F. Quatieri, A. C. Lammert, J. R. Williamson, Y. Yunusova, E. Godoy, D. D. Mehta, and J. R. Green, "Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis," in *Proc. INTERSPEECH 2016*. ISCA, 2016, pp. 1205–1209.
- [6] P. Östberg, N. Bogdanović, and L.-O. Wahlund, "Articulatory agility in cognitive decline," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 5, pp. 269–274, 2009.
- [7] S. Ntymenou, I. Tsantzali, T. Kalamatianos, K. I. Voumvourakis, E. Kapaki, G. Tsivgoulis, G. Stranjalis, and G. P. Paraskevas, "Blood biomarkers in frontotemporal dementia: review and meta-analysis," *Brain Sciences*, vol. 11, no. 2, p. 244, 2021.
- [8] D. Banerjee, A. Muralidharan, A. R. H. Mohammed, and B. H. Malik, "Neuroimaging in dementia: a brief review," *Cureus*, vol. 12, no. 6, 2020.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.
- [11] S. de la Fuente García, C. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.
- [12] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic Alzheimer's disease detection from speech and language," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1784–1797, 2020.
- [13] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra, "Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks," *Journal of clinical and experimental neuropsychology*, vol. 40, no. 9, pp. 917–939, 2018.
- [14] Y. Pan, V. S. Nallanthighal, D. Blackburn, H. Christensen, and A. Härmä, "Multi-task estimation of age and cognitive decline from speech," in *Proc. ICASSP 2021*, 2021, pp. 7258–7262.
- [15] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification," in *Proc. INTERSPEECH 2020*, 2020, pp. 4806–4810.
- [16] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting AD," in *Proc. INTERSPEECH 2019*, 2019, pp. 4105–4109.
- [17] B. Mirheidari, D. Blackburn, R. O'Malley, A. Venneri, T. Walker, M. Reuber, and H. Christensen, "Improving cognitive impairment classification by generative neural network-based feature augmentation," in *Proc. INTERSPEECH 2020*, 2020, pp. 2527–2531.
- [18] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Toward the automation of diagnostic conversation analysis in patients with memory complaints," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 373–387, 2017.
- [19] T. T. W. T. A. A. H. Alhanai, "Detecting cognitive impairment from spoken language," Ph.D. dissertation, Massachusetts Institute of Technology, 2019.
- [20] N. Dawalatabad, Y. Gong, S. Khurana, R. Au, and J. Glass, "Detecting dementia from long neuropsychological interviews," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 5270–5283.
- [21] L. Matošević and A. Jović, "Accurate detection of dementia from speech transcripts using roberta model," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2022, pp. 1478–1484.
- [22] Z. Liu, L. Proctor, P. Collier, D. Casenhiser, E. J. Paek, S. O. Yoon, and X. Zhao, "Machine learning of transcripts and audio recordings of spontaneous speech for diagnosis of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 17, p. e057556, 2021.
- [23] B. Mirheidari, D. Blackburn, and H. Christensen, "Automatic cognitive assessment: Combining sparse datasets with disparate cognitive scores," in *Interspeech*, 2022.
- [24] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *Proc. INTERSPEECH 2020*, 2020, pp. 4961–4965.
- [25] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," *Computer Speech and Language*, vol. 53, pp. 65–79, 2019.
- [26] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-Based Alzheimer's dementia detection through spontaneous speech," in *Proc. INTERSPEECH 2021*, 2021, pp. 3810–3814.
- [27] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "Wavbert: Exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection," in *Proc. INTERSPEECH 2021*, 2021, pp. 3790–3794.
- [28] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for Alzheimer's dementia through spontaneous speech," in *Proc. INTERSPEECH 2020*, 2020.
- [29] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in *Proc. INTERSPEECH 2020*, 2020, pp. 2162–2166.
- [30] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.
- [31] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.
- [32] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velazquez, P. Zelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios," in *Proc. INTERSPEECH 2021*, 2021.
- [33] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," *arXiv preprint arXiv:2104.09356*, 2021.
- [34] L. Gauder, L. Pepino, L. Ferrer, and P. Riera, "Alzheimer disease recognition using speech-based embeddings from pre-trained models," in *Proc. INTERSPEECH 2021*, 2021, pp. 3795–3799.
- [35] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for Alzheimer's disease detection," *arXiv preprint arXiv:2106.01555*, 2021.
- [36] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," *arXiv preprint arXiv:2004.06833*, 2020.
- [37] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. M. Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *Proc. INTERSPEECH 2020*, 2020, pp. 2182–2186.
- [38] A. Mittal, S. Sahoo, A. Datar, J. Kadiwala, H. Shalu, and J. Mathew, "Multi-modal detection of Alzheimer's disease from speech and text," *arXiv preprint arXiv:2012.00096*, 2020.
- [39] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," *arXiv preprint arXiv:2106.09668*, 2021.
- [40] A. Adhikari, "DocBERT: BERT for document classification," *arXiv preprint arXiv:1904.08398*, 2019.
- [41] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.
- [42] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," in *Proc. INTERSPEECH 2020*, 2020.
- [43] Luz, Saturnino and Haider, Fasih and de la Fuente, Sofia and Fromm, Davida and MacWhinney, Brian, "Detecting cognitive decline using speech only: The ADReSSo challenge," *medRxiv*, 2021.

- [44] T. Searle, Z. Ibrahim, and R. Dobson, “Comparing natural language processing techniques for Alzheimer’s dementia prediction in spontaneous speech,” *arXiv preprint arXiv:2006.07358*, 2020.
- [45] A. Pompili, T. Rolland, and A. Abad, “The INESC-ID multi-modal system for the ADReSS 2020 challenge,” *arXiv preprint arXiv:2005.14646*, 2020.
- [46] S. Farzana and N. Parde, “Exploring MMSE score prediction using verbal and non-verbal cues,” in *Proc. INTERSPEECH 2020*, 2020, pp. 2207–2211.
- [47] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, “Exploiting multi-modal features from pre-trained networks for Alzheimer’s dementia recognition,” *arXiv preprint arXiv:2009.04070*, 2020.
- [48] B. Mirheidari, D. Blackburn, and H. Christensen, “Automatic cognitive assessment: Combining sparse datasets with disparate cognitive scores,” *Proc. INTERSPEECH 2022*, pp. 2463–2467, 2022.
- [49] D. A. Snowdon, “Healthy aging and dementia: findings from the nun study,” *Annals of internal medicine*, vol. 139, no. 5\_Part\_2, pp. 450–454, 2003.
- [50] G. W. Ross, J. L. Cummings, and D. F. Benson, “Speech and language alterations in dementia syndromes: Characteristics and treatment,” *Aphasiology*, vol. 4, no. 4, pp. 339–352, 1990.
- [51] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [52] A. De Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
- [53] S. Luz, S. de la Fuente, and P. Albert, “A method for analysis of patient speech in dialogue for dementia detection,” *arXiv preprint arXiv:1811.09919*, 2018.
- [54] S. Luz, “Locating case discussion segments in recorded medical team meetings,” in *Proceedings of the third workshop on Searching spontaneous conversational speech*, 2009, pp. 21–30.
- [55] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, “Pauses for detection of alzheimer’s disease,” *Frontiers in Computer Science*, vol. 2, p. 624488, 2021.
- [56] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, “Using state of the art speaker recognition and natural language processing technologies to detect alzheimer’s disease and assess its severity,” in *Interspeech*, 2020, pp. 2177–2181.
- [57] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, “Multilingual Alzheimer’s dementia recognition through spontaneous speech: a signal processing grand challenge,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.
- [58] L. Jin, Y. Oh, H. Kim, H. Jung, H. J. Jon, J. E. Shin, and E. Y. Kim, “Consen: Complementary and simultaneous ensemble for alzheimer’s disease detection and mmse score prediction,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.
- [59] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
- [60] E. Edwards, C. Dognin, B. Bollepalli, M. Singh, and V. Analytics, “Multiscale system for Alzheimer’s dementia recognition through spontaneous speech,” in *Proc. INTERSPEECH 2020*. ISCA, 2020, pp. 2197–2201.
- [61] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic features identify alzheimer’s disease in narrative speech,” *Journal of Alzheimer’s Disease*, vol. 49, no. 2, pp. 407–422, 2016.
- [62] M. Yancheva and F. Rudzicz, “Vector-space topic models for detecting Alzheimer’s disease,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.
- [63] V. Masrani, “Detecting dementia from written and spoken language,” Ph.D. dissertation, University of British Columbia, 2018.
- [64] U. Sarawgi, W. Zufikar, N. Soliman, and P. Maes, “Multimodal inductive transfer learning for detection of Alzheimer’s dementia and its severity,” *arXiv preprint arXiv:2009.00700*, 2020.
- [65] M. Rohanian, J. Hough, and M. Purver, “Alzheimer’s dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs,” *arXiv preprint arXiv:2106.15684*, 2021.
- [66] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, “Learning alignment for multimodal emotion recognition from speech,” *arXiv preprint arXiv:1909.05645*, 2019.
- [67] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.
- [68] B. Sharma, M. Madhavi, and H. Li, “Leveraging acoustic and linguistic embeddings from pretrained speech and language models for intent classification,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7498–7502.
- [69] H. Goodglass and E. Kaplan, *The assessment of aphasia and related disorders*. Lea & Febiger, 1972.
- [70] Y. Pan, “Linguistic- and acoustic-based automatic dementia detection using deep learning methods,” Ph.D. dissertation, University of Sheffield, 2022.
- [71] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, “Dementia detection using automatic analysis of conversations,” *Computer Speech & Language*, vol. 53, pp. 65–79, 2019.
- [72] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” IEEE Signal Processing Society, Tech. Rep., 2011.
- [73] V. Manohar, D. Povey, and S. Khudanpur, “JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning,” in *2017 IEEE automatic speech recognition and understanding workshop (ASRU)*, 2017, pp. 346–352.
- [74] Y. Zhang, Y. Pan, J. Zhang, X. Ji, Y. Zhang, Lu, and Mingyu, “Advancing alzheimer’s identification under noise conditions: A two-step framework utilizing self-supervised pretraining,” *Applied Acoustics*, 2024 (Under review).
- [75] Y. Pan, Y. Shi, H. Zhang, and M. Lu, “Swin-bert: A feature fusion system designed for speech-based alzheimer’s dementia detection,” *IEEE Signal Processing Letters*, 2024 (Under review).
- [76] Y. Pan, M. Lu, Y. Shi, and H. Zhang, “A path signature approach for speech-based dementia detection,” *IEEE Signal Processing Letters*, 2023.
- [77] F. Haider, S. De La Fuente, and S. Luz, “An assessment of paralinguistic acoustic features for detection of Alzheimer’s dementia in spontaneous speech,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.
- [78] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, “CNN architectures for large-scale audio classification,” in *ICASSP 2017-2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.
- [79] L. Ilias, D. Askounis, and J. Psarras, “Detecting dementia from speech and transcripts using transformers,” *Computer Speech & Language*, vol. 79, p. 101485, 2023.
- [80] S. Hollands, D. Blackburn, and H. Christensen, “Evaluating the performance of state-of-the-art asr systems on non-native english using corpora with extensive language background variation,” 2022.



**Yilin Pan** received the M.Sc. and Ph.D. degrees from the Harbin Institute of Technology, China and the University of Sheffield, UK, in 2017 and 2022, respectively. She is a Lecturer with the College of Artificial Intelligence at the Dalian Maritime University. Her main research interests are in the areas of pathological speech analysis, disordered speech recognition, speaker verification, and natural language processing.



**Heidi Christensen** received the M.Sc. and Ph.D. degrees from Aalborg University, Denmark, in 1996 and 2002, respectively. She is a Professor in Spoken Language Technologies in the Computer Science department at the University of Sheffield. Before that she held post-doc positions at the University of Sheffield, IDIAP, Switzerland and Aalborg University, Denmark. Her main research interests are in the areas of recognition of disordered speech, automatic processing of conversations, and the automatic detection and tracking of paralinguistic information such as emotions and general interactional behaviours.



**Bahman Mirheidari** received the M.Sc. and Ph.D. degrees from Shahid Beheshti University, Tehran, Iran and the University of Sheffield, UK, in 2003 and 2018, respectively. He is a Research Associate in the Computer Science department at the University of Sheffield. His main research interests are in the areas of medical applications of speech technology, automatic speech recognition of conversations, and emotion detection in conversations.