A Semi-Supervised Framework for Speech Confidence Detection using Whisper

Adam Wynn* and Jingyun Wang*
*Department of Computer Science, Durham, United Kingdom, DH1 3DF

***Abstract*—Automatic detection of speaker confidence is critical for adaptive computing but remains constrained by limited labelled data and the subjectivity of affective annotations. This paper proposes a semi-supervised hybrid framework that fuses deep semantic embeddings from the Whisper encoder with an interpretable acoustic feature vector composed of eGeMAPS descriptors and auxiliary probability estimates of vocal stress and disfluency. To mitigate reliance on scarce ground truth data, we introduce an Uncertainty-Aware Pseudo-Labelling strategy where a model generates labels for unlabelled data, retaining only high-quality samples for training. Experimental results demonstrate that the proposed approach achieves a Macro-F1 score of 0.751, outperforming self-supervised baselines, including WavLM, HuBERT, and Wav2Vec 2.0. The hybrid architecture also surpasses the unimodal Whisper baseline, yielding a 3% improvement in the minority class, confirming that explicit prosodic and auxiliary features provide necessary corrective signals which are otherwise lost in deep semantic representations. Ablation studies further show that a curated set of high confidence pseudo-labels outperforms indiscriminate large scale augmentation, confirming that data quality outweighs quantity for perceived confidence detection.**

***Index Terms*—Speech Confidence Detection, Computational Paralinguistics, Semi-Supervised Learning, Pseudo-Labelling, Whisper, Disfluency Detection.**

I. Introduction

EFFECTIVE communication is fundamental to human social interaction, enabling us to share knowledge, emotion, and intent [1]. Central to these interactions, speaker confidence serves as an important affective cue that directly influences listener perceptions of credibility, persuasiveness, and competence [2], [3]. Therefore, the automatic modelling of confidence presents significant opportunities within the field of affective computing. For example, agents capable of detecting confidence can provide adaptive scaffolding to support self-efficacy in learning environments [4]. Furthermore, the implications of confidence detection extend to mental health monitoring and empathetic Human-Computer Interaction (HCI), as low confidence is a known precursor to anxiety and social withdrawal [5].

In the context of affective computing, detecting confidence is a challenging task as it involves analysing both the semantic content and vocal delivery. A primary difficulty lies in resolving verbal and non-verbal discrepancies, situations where the text and audio features conflict. For example, a speaker may use assertive words, but their prosody indicates uncertainty through hesitation or pitch instability. Accurately identifying these discrepancies is essential for reliable confidence detection. Prior research has established that acoustic features such as pitch instability (jitter), intensity, dynamics, and speech rate vary with confidence levels [6]. However, automating this detection remains an open challenge. Most existing approaches have relied on small, manually annotated datasets and handcrafted features [7], which limits scalability. Furthermore, modern end-to-end approaches using deep speech representations (e.g., Wav2Vec 2.0, HuBERT) often prioritise lexical or phonemic fidelity over subtle prosodic dynamics [8], potentially failing to capture the subtle nuances of features that correlate with a lack of confidence.

This limitation is exacerbated by the scarcity of labelled data. Unlike basic discrete emotions, such as happiness, sadness and anger, which benefit from large-scale benchmarks, confidence datasets are small, subjective, and difficult to access. To address this, in this paper, a refined speech confidence dataset is constructed by re-annotating subsets of existing corpora including TED-LIUM [9], SEP-28K [10], CMU-MOSI [11] and MLCommons People's Speech [12]. This approach ensures coverage across diverse speaking styles and demographics whilst providing high-quality ground truth labels. However, since large-scale manual annotation is time-consuming, and to mitigate the reliance on a limited amount of annotated examples, a semi-supervised pseudo-labelling framework is employed. By leveraging a model to automatically label a larger unlabelled corpus, the training distribution is significantly expanded without incurring the cost of additional manual labelling.

This paper extends our preliminary work [13] regarding semi-supervised confidence detection, where the integration of neural embeddings from the Whisper encoder [14] with a set of handcrafted features was previously explored. Whisper was prioritised over acoustic self-supervised learning models such as Wav2Vec 2.0 [15] due to its massive weakly-supervised pre-training, which yields representations that are semantically richer and more robust to speaker variability. Whilst the initial framework demonstrated the potential of pseudo-labelling [16] to mitigate data scarcity, this paper more rigorously evaluates the necessity of both the architecture and the learning strategy, and introduces a more diverse and expanded ground-truth dataset with increased annotator reliability. Moreover, this work incorporates an expanded feature vector consisting of eGeMAPS functionals [17] along with disfluency and stress

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.
Manuscript received April 19, 2021; revised August 16, 2021.

auxiliary models to generate more robust pseudo-labels, and employs a Late Fusion strategy to integrate semantic and acoustic features without compromising the robustness of the semantic representation.

Therefore, a robust hybrid framework is proposed which fuses deep semantic embeddings from Whisper with a set of interpretable acoustic descriptors. Furthermore, to address data scarcity, an uncertainty-aware pseudo-labelling strategy that expands the training set by filtering for samples with the highest model certainty is introduced. This approach ensures that only high-quality samples are seen by the model, demonstrating that a small, curated curriculum of augmented data is significantly more effective than indiscriminate large-scale supervision.

Through this work, we address the following research questions:

RQ1: To what extent does the augmentation of training data with pseudo-labels yield a significant improvement in performance compared to a baseline trained exclusively on ground truth data?

RQ2: To what extent does prioritising data quality via uncertainty-aware filtering outweigh data quantity compared to standard pseudo-labelling in mitigating class imbalance and improving model robustness?

RQ3: To what extent does fusing explicit paralinguistic cues with deep semantic embeddings (Whisper) improve the detection of uncertainty compared to Whisper-only baselines?

In this paper, we advance the modelling of confidence through the following contributions. First, a hybrid framework is introduced which combines the semantic capabilities of Whisper embeddings with interpretable, handcrafted prosodic features. By employing a late-fusion strategy, this framework captures nuances of uncertainty often missed by deep encoders alone. Second, an Uncertainty-Aware Data Strategy is introduced to prioritise label quality over quantity. By filtering pseudo-labels based on labeller confidence, it is shown that the addition of a curated set of high-quality samples significantly improves generalisation compared to noisy large-scale augmentation. This research contributes to the broader field of affective computing by establishing one of the first dedicated frameworks for the automatic detection of speaker confidence.

II. Related Work

A. Confidence and Disfluency Detection

Confidence in speech is a complex paralinguistic construct perceived through a combination of vocal characteristics including pitch, dynamics, intensity, and articulation rate. Specifically, in this paper, confidence is defined as perceived speaker confidence rather than the speaker's internal state of self-efficacy or physiological stress. From a theoretical perspective, confidence maps closely to the Dominance dimension of the Valence-Arousal-Dominance (VAD) model [18]. Whilst many basic emotions (e.g., happiness or sadness) are distinguished primarily by valence and arousal [19], confidence is fundamentally characterised by high Dominance and Control [20]. This creates a significant ambiguity in the signal space as acoustic markers of dominance frequently overlap with those of emotions such as anger, making it difficult to distinguish confidence from other high-arousal states based on prosody alone.

To identify these specific signals, Jiang and Pell [6], [21] mapped the primary vocal characteristics of confidence. Their research demonstrated that confident speech is characterised by a higher f_0 (pitch) range and mean amplitude, whilst unconfident speech exhibits a reduced speaking rate, lower mean f_0, and frequent pauses.

Despite this theoretical grounding, the transition to automated detection has been gradual. Early systems focused primarily on general public speaking skills rather than the specific affective state of confidence. For instance, Trinh et al. [22] developed RoboCop, an automated coaching system that provides feedback on speech quality metrics such as pacing and volume. While effective for general performance scoring, such systems often rely on heuristic rules rather than learning the latent representations of confidence.

More recent approaches have adopted deep learning to model perceived confidence. Chanda et al. [23] proposed a deep audiovisual framework using Bi-Directional LSTMs (Bi-LSTM) to classify confidence into three levels (High, Medium, Low) from interview recordings. Their work demonstrated that fusing audio and visual modalities yields measurable performance gains over unimodal baselines. However, their reliance on a small, specific dataset (34 candidates) and older recurrent architectures limits the generalisability of their findings to in-the-wild scenarios. Similarly, Nair et al. [8] achieved 86.3% accuracy using a CNN trained on MFCCs, but the reliance on a small, private dataset further restricts scalability. Our work addresses these gaps by employing a semi-supervised framework anchored by the Whisper encoder which improves generalisability compared to previous methods, enabling robust confidence detection across different speakers despite the scarcity of labelled data.

Confidence also heavily influences speech fluency. A lack of confidence is strongly correlated with an increase in disfluencies, such as interjections and prolongations [24]. Consequently, disfluency detection serves as a critical auxiliary task. Current approaches to detecting speech disfluency heavily rely on annotated data, which is limited in availability. Among the existing datasets are SEP-28K [10], collected from online podcasts, and Fluencybank [25], a dataset for the study of fluency development of native and non-native adults and children. Several approaches for disfluency classification have emerged including passing spectrograms into residual networks followed by a bi-directional LSTM for stutter classification [26], or using Transcription Based Methods [27]. Still, these approaches require large labelled datasets.

To address the challenge of limited data, Mohapatra et al. [28] proposed DisfluencyNet, a Wav2Vec 2.0-based model with convolutional and fully connected layers, achieving over 5% improvement in disfluency detection compared to baselines with only a few minutes of data. Similarly, Liu et al. [29] added a Transformer layer to enhance Wav2Vec 2.0's generalisation for detecting disfluencies across languages. Ameer et al. [30] used the Whisper model for multi-class disfluency

classification, introducing an encoder-freezing strategy and refining the SEP-28K dataset. They demonstrated that Whisper outperformed Wav2Vec 2.0 in multi-class disfluency tasks, achieving an average F1 of 0.81, offering better generalisation and faster inference. However, their approach struggled to accurately classify fluent speech (F1=0.23) suggesting that multi-class labels introduce significant noise. Therefore, in this paper, disfluency detection is treated as a binary-classification task using the Whisper encoder, providing a useful auxiliary signal for detecting reduced speaker confidence.

B. Speech Feature Extraction and Speech Emotion Recognition

Confidence detection shares significant methodological overlap with Speech Emotion Recognition (SER) as both tasks rely on disentangling prosodic affect from linguistic content. Traditional SER approaches use handcrafted features such as Mel Frequency Cepstral Coefficients (MFCCs) [31], which capture the spectral envelope but often fail to model long-term temporal dynamics [32].

To capture these missing details, modern systems [33] have adopted the eGeMAPS [17] feature set. Unlike MFCCs, which only describe the general shape of the sound, eGeMAPS encodes specific, meaningful vocal patterns such as voice shakiness (jitter) and volume changes which are close to what listeners rely on to gauge a speaker's perceived confidence. Despite their interpretability, these features often struggle to generalise across diverse acoustic environments or capture the high-level semantic context of an utterance.

To address these limitations, the introduction of Self-Supervised Learning allows models to automatically extract rich, high-dimensional representations from audio. Pepino et al. [34] demonstrated that Wav2Vec 2.0 embeddings, when fused with prosodic features, outperform handcrafted baselines in SER. However, Wagner et al. [35] caution that transformer-based models can become over-reliant on linguistic information, potentially ignoring paralinguistic cues when the text is semantically dominant. This semantic bias is particularly relevant for confidence detection, where a speaker may say something with high semantic confidence but with an unconfident-sounding voice. To address this, recent studies have explored the weakly-supervised Whisper model for paralinguistic tasks. Goron et al. [36] and Osman et al. [37] found that Whisper embeddings generalise better to unseen speakers than Wav2Vec 2.0 or HuBERT, likely due to the massive diversity of its weak-supervision training data.

However, because ASR models like Whisper are optimised to minimise Word Error Rate, they naturally prioritise linguistic invariance over prosodic variability. Consequently, fine-grained cues like pitch instability may be treated as secondary to the lexical content. Our hybrid approach complements the strong semantic backbone of Whisper by explicitly reintroducing these paralinguistic signals by fusing explicit paralinguistic cues with the Whisper encoder embeddings.

C. Semi-Supervised Learning in Affective Computing

The scarcity of labelled data necessitates data-efficient learning strategies. Whilst Transfer Learning from large pre-trained models (e.g., Whisper, Wav2Vec 2.0) provides strong initial representations, adapting these models to specific downstream tasks typically requires fine-tuning on labelled data, which remains a bottleneck. To address this, semi-supervised learning techniques such as Pseudo-Labelling [16] leverage a model to generate target labels for unlabelled data, expanding the overall amount of available data. However, standard pseudo-labelling is prone to confirmation bias, where the hybrid model overfits to the labellers' incorrect predictions [38]. This risk is exacerbated in affective computing, where ground-truth labels are subjective and class boundaries are ambiguous.

To mitigate this, recent literature emphasises uncertainty-aware filtering. Approaches such as FixMatch [39] use strict confidence thresholds to ensure that only high-quality samples contribute to the model. In the speech domain, this approach is often extended via Cross-View Supervision, where models trained on different 'views' of the data supervise each other. We adopt this philosophy by using a separate pseudo-labeller model trained only on acoustic features such as eGeMAPS, rather than allowing the Whisper model to teach itself. This prevents the system from reinforcing its own semantic biases. By generating labels based on explicit vocal cues, we create a high-precision dataset that is robust to the noise often found in real-world audio.

III. METHODOLOGY

Motivated by recent progress in deep learning approaches to audio representation and classification, this study proposes a hybrid framework for speech confidence detection. As illustrated in Fig. 1, the method integrates interpretable prosodic and spectral features with deep audio embeddings derived from the Whisper-base encoder. The entire training and evaluation protocol is executed via a robust 5-Fold Cross-Validation Pipeline which comprises five key stages: (A) Dataset Creation and Annotation, (B) Feature Vector Creation, (C) Auxiliary Disfluency and Stress Detection, (D) Model-based pseudo-label generation, and (E) Hybrid model training for confidence detection which incorporates information from both the feature vector and deep audio embeddings effectively to develop a robust system for classifying speaker confidence levels on limited annotated data.

The pipeline begins with the creation of a small annotated dataset labelled by human raters. To address the challenge of limited available labelled data, a pseudo-labelling approach is used to expand the dataset sufficiently to train the model effectively.

A. Dataset and Audio Preprocessing

Due to the absence of publicly available datasets and standard benchmarks for perceived confidence, a manual confidence annotation process was conducted. A custom dataset (D_L) comprising $N = 600$ utterances between 5 and 12 seconds was curated by sampling from the TED-LIUM [9], CMU-MOSI [11], MLCommons People's Speech [12], and SEP-28K [10] datasets, supplemented with additional recordings

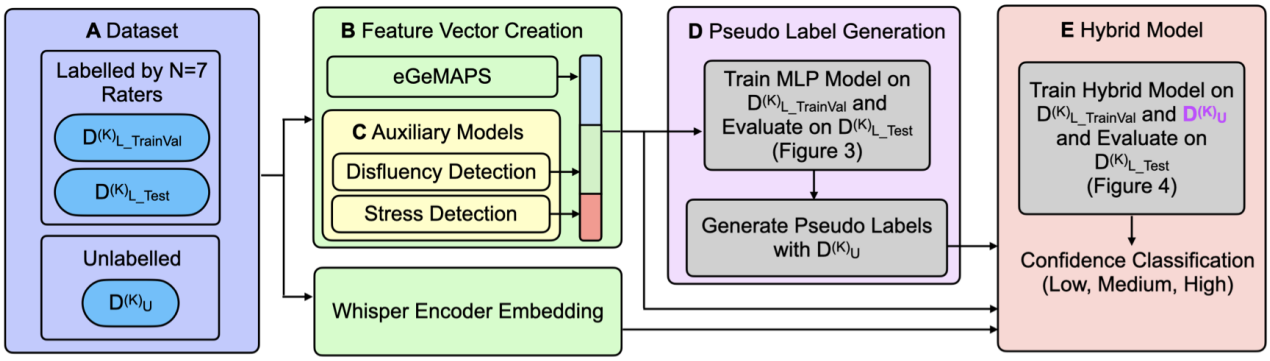

Fig. 1. Overall Pipeline of the Confidence Classification System during Training for Fold k

TABLE I
INTER-RATER RELIABILITY ANALYSIS ($N = 576$, $k = 7$ RATERS)

Measure	ICC	95% CI	F	df1	df2	p-value
Single Rater	0.48	[0.44, 0.53]	8.21	575	3450	< 0.001
Average (Consensus)	**0.87**	[0.85, 0.89]	8.21	575	3450	$\mathbf{< 0.001}$

Model: Two-way random effects, absolute agreement (ICC 2,k).

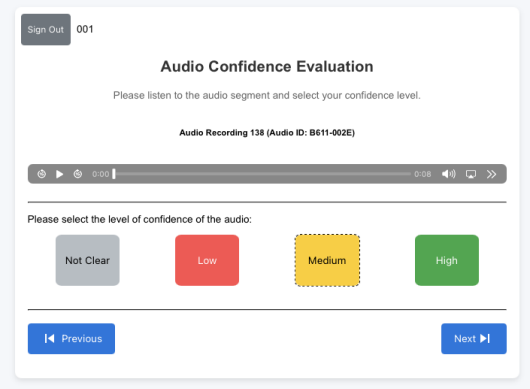

Fig. 2. User Interface of the Labelling System

of both native and non-native English speakers in beginner speech contests to ensure diversity.

As confidence in speech is subjective, to ensure robust labels, each clip was independently annotated by 7 English-fluent annotators (4 male, 3 female) using a three-level ordinal scale: *low*, *medium*, and *high*. We computed the Intraclass Correlation Coefficient (ICC) using a two-way random-effects model. Analysis was performed on $N = 576$ samples. The remaining 24 samples were excluded from this specific calculation because one or more raters flagged the audio as 'Not Clear', resulting in incomplete ordinal data for those instances. However, these samples were retained in the final dataset ($N = 600$) as a clear consensus label could still be derived from the remaining valid annotations. Whilst single-rater agreement was moderate ($ICC = 0.48$), reflecting individual subjectivity, the average-measures reliability was excellent ($ICC = 0.87$), confirming that the collective consensus provides a stable training signal. Individual labels were aggregated using the Dawid-Skene (DS) probabilistic model [40], which further refines the ground truth by weighting annotators based on their consistency. The final curated dataset comprises 300 high, 210 medium, and 90 low-confidence clips.

To overcome the limitation of the dataset's small size and inherent class imbalance, a model-based pseudo-labelling approach is applied to expand it. The augmented dataset, sampled from the same corpora as D_L, with all segments appearing in D_L excluded, is then used for both training and validation (development). Further details on this technique are provided in Section III-D. All audio files are converted to WAV format, resampled to 16 kHz, and converted to mono. Background noise is removed from each clip using the noisereduce Python library [41].

As previously mentioned, to ensure rigorous evaluation without data leakage, the entire pipeline operates under a consistent stratified 5-fold cross-validation framework. The dataset D_L is partitioned into 5 folds once and these splits remain fixed throughout the experiment. The training folds used to build the pseudo-label model ($D^{(k)}_{L\_TrainVal}$) are the exact same folds used to train the final hybrid model, and the held-out test sets ($D^{(k)}_{L\_Test}$) are identical for both evaluations. Importantly, for any given fold k, the held-out test set $D^{(k)}_{L\_Test}$ is never seen by either model during training.

B. Feature Vector Creation

For every input audio x_i, two distinct modalities are extracted. (1) Whisper Base Encoder: The Whisper Base encoder is used to extract 512-dimensional semantic embeddings. The first three encoder layers are frozen to preserve linguistic representations. (2) Feature Vector: We construct a 94-dimensional acoustic vector $\mathbf{f}_i$ containing eGeMAPS prosodic features (88 dim) and Auxiliary Scores (6 dim) for Disfluency and Stress.

1) Acoustic and Prosodic Features (OpenSMILE): To capture paralinguistic cues, we used the OpenSMILE toolkit [42] to extract the eGeMAPSv02 (extended Geneva Minimalistic Acoustic Parameter Set) [17]. This standardised feature set consists of 88 functionals derived from low-level descriptors across frequency, energy, spectral and temporal domains.

These features provide a robust, interpretable baseline for detecting perceived confidence, and are used as inputs for the pseudo-labelling models.

2) Auxiliary High-Level Features: To augment the low-level acoustic descriptors, we append 6 high-level probability scores derived from auxiliary Whisper-based classifiers (Section III-C):

$$\mathbf{f}_{aux} = [\mathbf{d}_i, s_i] \in \mathbb{R}^6$$

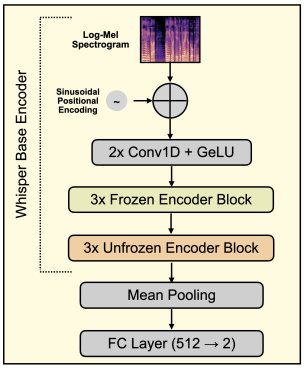

Fig. 3. Architecture of Auxiliary Models - Disfluency and Stress Classifiers

where:

- $\mathbf{d}_i \in \mathbb{R}^5$ represents the calibrated probabilities for five distinct disfluency types: Block, Prolongation, Interjection, Word Repetition, and Sound Repetition.
- $s_i \in \mathbb{R}^1$ represents the calibrated probability of perceived vocal stress.

The final feature vector $\mathbf{f}_i = [\text{eGeMAPS}_i; \mathbf{d}_i; s_i]$ is z-score normalised using statistics from the training partition to ensure numerical stability during training.

C. Auxiliary Disfluency and Stress Detection

1) Disfluency Detection: Confidence can also influence the frequency of speech disfluencies, as speakers who feel less confident often exhibit more pauses, repetitions, and filler words [24]. Therefore, as disfluency patterns can serve as an indicator of confidence, detecting these disfluencies is an essential step in confidence detection.

a) Dataset: Among existing English-language resources for disfluency detection, two primary datasets are widely used: SEP-28K [10], which consists of 28,000 three-second clips extracted from eight podcasts, and FluencyBank [25], which contains approximately 4,000 instances designed for the study of fluency development in native and non-native adults and children. Each sample in these datasets was annotated for the presence of the disfluency types in table II.

TABLE II
DISFLUENCY TYPES INCLUDED IN SEP-28K AND FLUENCYBANK.

Type	Definition	Example
Prolongations	Elongation of speech sounds, often reflecting hesitation or planning.	"ssssso"
Blocks	Temporary stoppage of airflow, producing audible silence before the next phoneme.	(pause)
Sound Repetitions	Repetition of individual phonemes or syllables.	"b-b-but"
Word Repetitions	Repetition of entire words, indicating hesitation or correction.	"I I think"
Interjections	Filler or non-lexical utterances that signal hesitation or low confidence.	"uh," "um," "like"

The SEP-28K-Extended (SEP-28K-E) variant [43] was introduced to mitigate speaker imbalance in the original dataset, which was dominated by four main podcast hosts. It separates training and evaluation speakers to improve cross-speaker generalisation. The SEP-28K-E-Merged version further integrates FluencyBank data for testing, enabling cross-corpus evaluation and improved robustness to differences in recording conditions and speaking styles. This merged dataset has therefore become a benchmark for disfluency detection due to its size, diversity, and well-defined splits. Therefore, in this study, we employ the SEP-28K-E-Merged dataset [30].

b) Data Preprocessing: Each three-second audio clip in SEP-28K-E-Merged was labelled independently by three annotators so only the audio clips where the majority of raters agreed were used and data was excluded if any rater indicated they were unsure. Additionally, data was excluded which had poor audio quality, was difficult to understand, or contained music or non-speech content.

To examine the effect of class imbalance on disfluency detection, we train models using two fluent:disfluent sampling ratios: 0.8 and 1.0. Whilst the training set was balanced to prevent bias toward the fluent majority class, the validation and test sets retain their natural class distributions to enable realistic performance evaluation

c) Model Architecture: The model architecture (Fig. 3) uses HuggingFace's WhisperForAudioClassification implementation with the Whisper-base encoder [14] for binary disfluency detection. The Whisper-base encoder processes 80-dimensional log-mel spectrograms to produce contextualised acoustic representations. The first three encoder layers were frozen to preserve pre-trained linguistic representations. The classification head performs pooling over the encoder outputs and applies a linear transformation to produce two output logits for binary classification. Cross-entropy loss is used for optimisation during training with the AdamW optimiser [44] (learning rate 2.5×10^{-5}; weight decay 1×10^{-5}), and early stopping based on validation loss.

2) Stress Detection:

a) Dataset: To train and evaluate the stress classification model, three publicly available emotional speech corpora were combined: RAVDESS [45], SAVEE [46], and TESS [47]. Each dataset contains recordings of actors expressing a range of emotions, which were re-labelled into low stress and high stress, to better align with the goals of confidence detection.

Following the affective mapping protocols established in prior work [48], [49], in the RAVDESS dataset, emotions such as neutral, calm, and happy were labelled as low stress, while sad, angry, and surprised were labelled as high stress. The SAVEE dataset followed a similar mapping, with angry and sad labelled as high stress, and neutral and happy as low stress. In TESS, neutral and happy were considered low stress, and sad, angry, and surprised as high stress. After balancing, this results in 1460 low stress and 1460 high stress labels.

b) Data Preprocessing: All audio samples were converted into 16 kHz, single-channel waveforms and normalised prior to Mel-spectrogram extraction. The combined dataset provided a diverse set of emotional speech instances across multiple speakers and recording conditions, enhancing the model's robustness.

c) Model Architecture: The stress classification model follows a similar architecture to the disfluency model, us-

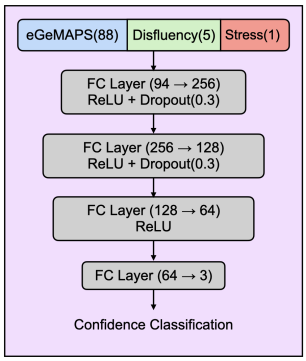

Fig. 4. Architecture of MLP Labeller

ing HuggingFace's WhisperForAudioClassification implementation to distinguish between low and high stress levels from 80-dimensional log-mel spectrograms.

However, due to the smaller number of samples per stress category compared to large disfluency corpora, the use of stratified 10-fold cross-validation with early stopping (patience of 10 epochs) was employed with 20% data left out for testing, to ensure robust performance evaluation across the binary classification problem.

3) Probability Calibration: To ensure the reliability of these features for the downstream task post-hoc temperature scaling [50] was applied. This involved dividing the logits $\mathbf{z}$ by a scalar parameter T before the softmax activation. This parameter T is optimised to minimise negative log-likelihood on the validation set, ensuring the output probabilities accurately reflect the model's true confidence.

D. Model-Based pseudo-Label Generation

To address the limitation of a small ground truth dataset ($N_{GT} = 600$), a model-based pseudo-labelling approach was used to generate additional labelled data by propagating labels from the ground truth to a larger unlabelled corpus. After the feature vectors ($\mathbf{f}_i$) were created, a multi-layer perceptron (MLP) was trained on the ground truth dataset feature vectors to classify speech confidence into three levels: low, medium, and high, as shown in Figure 4. The model was trained with Adam optimiser [51], cross-entropy loss, and a learning rate of 0.001. This process did not involve the audio or the Whisper embeddings directly but instead relied only on the feature vector representations ($\mathbf{f}_i$) of each sample in an attempt to prevent data leakage.

The trained MLP was then applied to the unlabelled corpus to generate probability distributions over the three confidence classes. To mitigate the risk of confirmation bias, where the main hybrid model learns the labeller's mistakes, we applied a strict confidence thresholding filter (τ). Only samples where the model's prediction confidence exceeded the threshold were retained as pseudo-labels. Based on empirical ablation studies, we initially set $\tau = 0.8$. This filtering removes the ambiguous confidence samples where the MLP was uncertain, curating a high-precision dataset with on average $N \approx 1194 \pm 345$ audio samples across each fold from an initial pool of 10589 unlabelled segments.

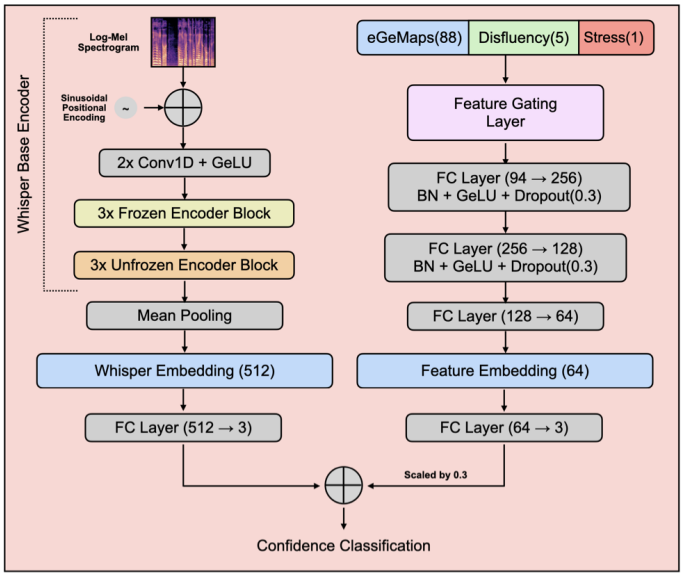

Fig. 5. Architecture of Hybrid Confidence Model

However, the filtering process naturally skews the distribution and low confidence samples were less frequent than High. To avoid physical downsampling in order to preserve the full diversity of the pseudo-labeled corpus, we used a Weighted Random Sampler that oversamples the minority classes ensuring that every training batch contains a uniform class distribution. The resulting pseudo-labelled dataset, combined with the original ground truth samples, provided a significantly larger dataset for subsequent stages of the model development pipeline, enabling improved generalisation and performance of the final confidence classification system.

E. Hybrid Model Training

The final stage of the pipeline involves training the hybrid model. As illustrated in Fig. 5, this model integrates the deep semantic capabilities of Whisper with the interpretable acoustic signals from the feature vector.

1) Architecture: The model operates via two parallel processing streams:

- Whisper Stream: The pre-processed audio is input into the Huggingface implementation of the Whisper-base encoder. The resulting 512-dimensional embeddings are passed through a linear projection head to generate semantic logits. To prevent overfitting and preserve pre-trained linguistic knowledge, the first three encoder layers are frozen.
- Feature Vector Stream: The 94-dimensional feature vector $\mathbf{f}_i$ is passed through a Feature Gating Layer which applies a learnable sigmoid mask to suppress irrelevant features before the vector is further processed by the MLP. As shown in Figure 5, the first two fully connected layers of the MLP are each followed by Batch Normalisation and GELU activation to stabilise training dynamics, along with a Dropout layer ($p = 0.3$) to prevent overfitting on the limited ground truth data.

To combine the two modalities, a late fusion strategy is used, where the final prediction is derived from a weighted sum of the logits. A scaling factor of $\lambda = 0.3$ is assigned to the feature vector so that Whisper provides the base confidence estimate, whilst the Feature Vector acts as a corrective signal.

2) Training Strategy: The model is trained on the union of the fold-specific ground truth and the pseudo-labelled dataset ($D^{(k)}_{L\_TrainVal} \cup D^{(k)}_{U}$). To ensure the model prioritises verified human annotations despite the volume of pseudo-labels, we employ a Source-Boosted Loss Function:

$$\mathcal{L} = \omega_{class} \cdot (\mathcal{L}_{CE}(y_L, \hat{y}) \cdot 18.0 + \mathcal{L}_{CE}(y_U, \hat{y})) \quad (1)$$

where the ground truth loss is scaled by a factor of 18. This scaling factor was determined empirically to normalise the gradient contribution of the two datasets. Additionally, a class weight $\omega_{med} = 1.2$ ($\omega_{low,high} = 1.0$) is applied to the minority Medium class to improve boundary separation. Optimisation is performed using AdamW [44] with a cosine annealing scheduler. A differential learning rate strategy is employed where the pre-trained Whisper stream is fine-tuned with a lower rate of 2.5×10^{-5} whilst the MLP feature stream is trained with a rate of 1×10^{-3}.

In our preliminary work [13], we additionally employed a complex downstream sequence model. However, this study demonstrates that pre-trained semantic features are sufficient for direct mapping. This simplification reduces trainable parameters while improving F1 score, highlighting the efficiency of the proposed model.

IV. Results

A. MLP Labeller

To generate pseudo-labels for the final hybrid model, a Multi-Layer Perceptron (MLP) was first trained on the engineered feature vector (f_i) only. For each outer fold k, the MLP was trained on the training partition of the ground truth dataset ($D^{(k)}_{L\_TrainVal}$) using an internal 5-fold cross-validation and its performance is reported on the held-out Test Set ($D^{(k)}_{L\_Test}$).

On the test sets ($D^{(k)}_{L\_Test}$) the MLP labeller achieved a mean macro F1-score of 0.746. The MLP demonstrated stronger capability in distinguishing low (F1=78.2) and high confidence (F1=81.8) but struggled more with the middle confidence class (F1=64.4), where misclassifications were primarily concentrated between adjacent classes (e.g., Medium misclassified as High, rather than Low misclassified as High). Therefore, as mentioned in section III-D a confidence threshold (τ) is applied to filter out the ambiguous labels where the MLP was less confident in its predictions, creating a high-precision dataset to supervise the downstream hybrid model.

B. Auxiliary Disfluency and Stress Detection Results

To validate the quality of the auxiliary features fed into the hybrid model, the performance of the disfluency and stress detectors are evaluated in this section.

1) Disfluency Detection Results: To evaluate the effectiveness of different design choices in the disfluency detection component, an ablation study optimising model capacity (Whisper-Base vs. Tiny), fine-tuning strategy (Frozen vs. Unfrozen encoder), and class balancing was conducted. All studies were conducted on the SEP-28-K-E-Merged Test Set.

As shown in Table III, the best-performing result was using the Whisper-Base model with frozen encoder layers and a balance ratio of 0.8. Among the five disfluency categories, interjections achieved the highest performance (F1 = 0.90), followed by sound repetitions (F1 = 0.81) and prolongations (F1 = 0.73). In contrast, blocks and word repetitions were more difficult to classify, reaching F1-scores below 0.65. This pattern aligns with prior findings that such disfluencies are often acoustically subtle or context-dependent, making them harder to detect automatically [28], [29]. Moreover, regarding class balance, the 0.8 ratio (mild imbalance) achieved the highest F1 (0.77), outperforming the strictly balanced (1.0) setup, suggesting that moderate balancing better preserves natural speech variability and improves robustness.

Across all disfluency types and balance ratios, as shown in table IV, the best performing model uses labels where 2+ raters agree yields higher F1-scores than when all 3 raters agree. Whilst some prior work on SEP-28K [28], [29] suggests that unanimous labels yield higher annotation quality, our results show that the reduction in dataset size associated with this harms model generalisation, particularly for low-frequency disfluencies such as blocks and prolongations.

2) Stress Detection Results: Table V presents the mean 10-fold cross-validation performance of the four Whisper-based stress classifiers using stratified 10-fold cross-validation. Results are reported strictly on the held-out test set. All configurations achieve similarly strong results, with F1-scores ranging from 0.936 to 0.942.

Given the negligible performance gap, we adopt Whisper-Base with frozen lower layers for the remainder of the experiments to ensure architectural consistency with the disfluency detector and reduce computational cost during deployment by avoiding unnecessary fine-tuning.

C. Hybrid Model Results

1) Experimental Setup: To ensure a robust evaluation of the proposed hybrid model, as explained in section III, a stratified 5-Fold Cross-Validation scheme was used. In each fold, the model was trained on a combination of the ground truth dataset ($D^{(k)}_{L\_TrainVal}$), and the pseudo-labelled augmented data ($D^{(k)}_{U}$) with evaluation restricted strictly to the held-out test partition ($D^{(k)}_{L\_test}$) for each fold k. Model selection was performed using the Validation Macro-F1 score, ensuring that the reported test metrics correspond to the epoch with the highest performance on the validation set.

2) Results: The proposed architecture is compared against two unimodal baselines: Feature Vector Only, Whisper Encoder Only and the Proposed hyrbid architecture combining the Whisper encoder embeddings with the feature vector.

As shown in Table VI, the Whisper-only baseline outperforms the Feature Vector Only model (0.736 vs. 0.665 Macro-F1). This performance gap is likely due to the fundamental

TABLE III
F1-SCORES FOR EACH DISFLUENCY TYPE UNDER DIFFERENT MODEL CONFIGURATIONS AND BALANCE RATIOS.

Type	Base Frozen		Base Unfrozen		Tiny Frozen		Tiny Unfrozen	
	0.8	1.0	0.8	1.0	0.8	1.0	0.8	1.0
Blocks	0.642	**0.664**	0.591	0.582	0.554	0.536	0.562	0.533
Interjections	**0.900**	**0.900**	0.892	0.884	0.883	0.881	0.873	0.862
Prolongations	**0.730**	0.685	0.657	0.690	0.672	0.638	0.617	0.652
Sound Repetitions	**0.813**	0.793	0.795	0.763	0.724	0.642	0.701	0.670
Word Repetitions	0.743	**0.760**	0.729	0.709	0.644	0.631	0.591	0.577
Mean	**0.766**	0.760	0.733	0.726	0.695	0.666	0.669	0.659

TABLE IV
F1-SCORES FOR WHISPER-BASE DISFLUENCY DETECTION UNDER DIFFERENT ANNOTATION AGREEMENT CONDITIONS.

	2+ Raters Agree		All 3 Raters Agree	
Disfluency Type	Base Frozen		Base Frozen	
	0.8	1.0	0.8	1.0
Block	0.642	**0.664**	0.289	0.274
Interjection	**0.900**	0.900	0.823	0.830
Prolongation	**0.730**	0.685	0.643	0.681
Sound Repetition	**0.813**	0.793	0.664	0.734
Word Repetition	0.743	**0.760**	0.717	0.740
Mean F1	**0.766**	0.760	0.627	0.652

TABLE V
STRESS DETECTION PERFORMANCE ACROSS WHISPER VARIANTS.

Model	Frozen	F1-Score
Whisper Base	Yes	0.9385
Whisper Base	No	**0.9423**
Whisper Tiny	Yes	0.9413
Whisper Tiny	No	0.9357

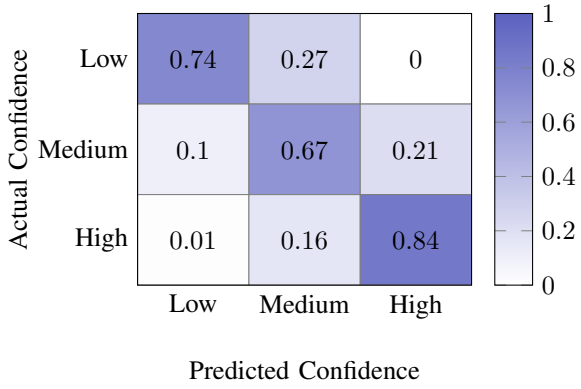

Fig. 6. Confusion Matrix for Hybrid Confidence Model

difference in capacity. The feature vector model is restricted to a limited set of 94 handcrafted descriptors whereas the Whisper encoder leverages deep, high-dimensional representations learned from 680k hours of diverse audio. Consequently, the Whisper embeddings likely capture a complex interplay of both phonetic and prosodic cues that simple feature engineering misses. Moreover, the feature vector model struggles most especially with the Medium Confidence class (0.532). This highlights the limitations of low-dimensional feature sets, which fail to capture the subtle dynamics required to distinguish confident from uncertain speech.

Despite the strong performance of the Whisper-Only baseline, the proposed hybrid architecture achieves the highest overall performance (0.751 Macro-F1), confirming that explicit acoustic supervision remains beneficial. Specifically, the hybrid model improves the minority class detection (low and medium confidence) compared to the Whisper baseline. This suggests that while Whisper implicitly captures prosody, it may miss subtle uncertainty markers (like jitter or hesitation) in favor of dominant linguistic patterns. The feature vector resolves this by emphasising these specific signals that the Whisper-Only baseline missed.

TABLE VI
MEAN MACRO-F1 SCORES OF THE HYBRID CONFIDENCE MODEL

Confidence	FV Only	Whisper Only	Proposed
Low	0.666 ± 0.098	0.714 ± 0.086	$\mathbf{0.744 \pm 0.068}$
Medium	0.532 ± 0.032	0.656 ± 0.080	$\mathbf{0.672 \pm 0.052}$
High	0.796 ± 0.032	$\mathbf{0.838 \pm 0.041}$	0.836 ± 0.036
Macro-F1	0.665 ± 0.041	0.736 ± 0.049	$\mathbf{0.751 \pm 0.041}$

Fig. 6 presents the confusion matrix for the proposed hybrid model. The results demonstrate that the model is highly robust against catastrophic errors. Specifically, the misclassification of Low confidence samples as High confidence is non-existent (0.00) on the test set, and less than 1% of High confidence samples are misclassified as Low confidence. Moreover, errors concentrate around the Medium confidence boundary, with 27% of Low confidence samples predicted as Medium. This confirms the model has learned the ordinal nature of confidence, restricting misclassifications to neighbouring classes rather than making random predictions.

To further analyse the model's decision-making, a t-SNE projection of the embeddings for the ground truth data was plotted using the ensemble model, which also confirms that the model has learned a meaningful ordinal manifold. As seen in the t-SNE plot (Fig. 7), the Low Confidence (Red) and High Confidence (Green) clusters occupy distinct regions of the latent space with minimal overlap. The Medium Confidence class (Yellow) acts as a transition between low and high confidence, validating its role as an ambiguous boundary class where linguistic and paralinguistic cues naturally overlap.

Moreover, to understand the specific acoustic drivers be-

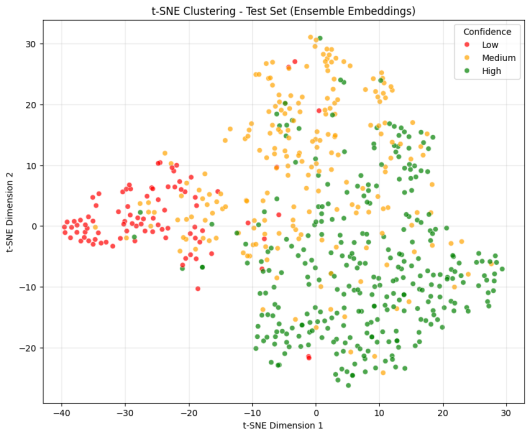

Fig. 7. t-SNE visualisation of test set embeddings using the ensemble model.

hind this clustering, the contribution of the features in the feature vector were analysed using SHAP (SHapley Additive exPlanations) [52]. As shown in figure 8, the Low Confidence predictions are driven heavily by spectral characteristics and vocal instability. Specifically, MFCC 3 Mean emerges as the top predictor, where low values strongly correlate with low confidence. MFCC 3 often relates to the quality of the vocal tract, where lower values result in a flatter shape often heard in mumbled speech. Jitter Standard Deviation and Jitter Mean, which are measures of pitch perturbation, are also amongst the most influential features, indicating that vocal tremor is a dominant signal for low confidence. Furthermore, disfluency types such as Prolongation and Block appear as strong positive predictors.

For the Medium Confidence class, stress is the top feature for this class, with distinct clustering suggesting that specific stress patterns likely serve as key differentiators for this ambiguous middle ground, where lower stress values make the model more likely to classify speech as medium. Interjection and Word Repetition also appear as top positive predictors, suggesting that the model identifies Medium confidence through indicators of hesitation rather than vocal quality (e.g. MFCC Mean and Jitter) that characterise Low confidence.

Finally, for the High Confidence class, the model primarily looks for the absence of disfluencies. The plot shows a strong negative relationship between disfluencies (eg. Word Repetition and Interjection) and high confidence where the presence of disfluencies makes the model less likely to predict high confidence. Consistent with the Low Confidence analysis, Jitter remains a top predictor, but exhibits an inverse relationship where the model strongly associates low jitter values with high confidence. Additionally, Pitch Range shows a positive correlation, where a wider pitch range, commonly associated with more confident speech [6], contributes to a high confidence classification, and in contrast, a higher pitch range makes a sample less likely to be classified as Medium confidence. However, the correlation between Stress and High Confidence could be because stress labels are generated by mapping emotions to stress, and the resulting features capture acoustic traits that overlap with confident speech.

D. Impact of Training Strategies

To assess the robustness of the semi-supervised pipeline, we conducted a series of studies regarding encoder adaptability and data filtering.

1) Impact of Unfreezing Whisper Encoder Blocks: Unfreezing all encoder blocks of Whisper-base resulted in a Macro-F1 of 0.742 with class-wise scores of 0.739 (Low), 0.658 (Medium), and 0.830 (High). Therefore, freezing the first 3 layers offered the optimal balance, achieving Macro-F1 of 0.751. This suggests that the initial layers contain fundamental acoustic filters that are best left preserved, whilst the higher-level layers require adaptation to capture the semantic nuances of confidence.

2) Impact of Semi-Supervised Data Strategy: To validate the semi-supervised learning framework, the impact of pseudo-label quality versus quantity is examined. The proposed uncertainty filtering approach ($\tau > 0.8$; $N \approx 1194$) is compared against two ablations: (1) training on ground truth data only ($N = 480$), and (2) using all generated pseudo-labels without filtering ($N = 11069$).

As shown in Fig. 9, training on the full, unfiltered pseudo-labelled (PL) dataset D_U caused performance to fall to a Macro-F1 of 0.685. This could suggest that the noise and bias from the pseudo-labeller outweighs the benefits of additional labelled data, highlighting that high data quality is more important than simply increasing the amount of data available. On the other hand, relying exclusively on the ground truth (GT) data yielded a Macro-F1 of 0.726. This suggests that the model might have overfitted to the limited ground truth data, limiting its capacity to generalise to unseen speakers.

The proposed uncertainty filtering approach achieves the optimal F1 (0.751 by leveraging the scale of pseudo-labels while maintaining quality through strict filtering. This strategy proves effective especially for the minority classes (Low Confidence and Medium Confidence detection), demonstrating that high-quality, diverse data is essential for capturing the subtle acoustic variations of uncertainty that are underrepresented in the small ground truth set.

3) Impact of Model Architecture: To isolate the contribution of the underlying representation, the proposed Whisper-Base encoder was benchmarked against three state-of-the-art Self-Supervised Learning models (Wav2Vec 2.0 [15], HuBERT [53], WavLM [54]) and a smaller architectural variant (Whisper-Tiny). Table VII summarises the results. Wav2Vec 2.0 yields the lowest performance (0.661 Macro-F1), but HuBERT and WavLM show notable gains, with WavLM achieving a strong score of 0.737. Notably, WavLM matches the proposed model in the Medium confidence class (0.672).

However, the proposed Whisper-Base architecture achieves the highest overall stability and performance (0.751 Macro-F1). Specifically in the Low Confidence class, Whisper-Base outperforms WavLM. This superior generalisation on the minority class could suggest that the massive pre-training is beneficial. WavLM is trained on approximately 94k hours of unlabelled data [54], whereas Whisper leverages 680k hours of weakly supervised audio, allowing the encoder to learn a far more robust feature space for difficult, ambiguous examples. Furthermore, increasing model capacity from Tiny (0.728)

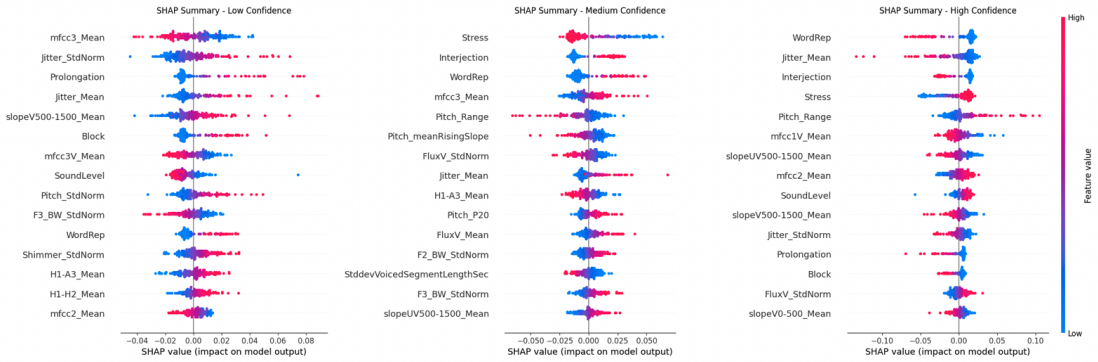

Fig. 8. SHAP Feature Importance Analysis for Low, Medium and High Confidence.

TABLE VII
COMPARISON OF DIFFERENT MODEL ARCHITECTURES.

Architecture	Wav2Vec 2.0	HuBERT	WavLM	Whisper-Tiny	Wynn et al, 2025 [13]	Whisper-Base (Proposed)
Low	0.604 ± 0.129	0.704 ± 0.068	0.726 ± 0.083	0.717 ± 0.089	0.628 ± 0.185	$\mathbf{0.744 \pm 0.068}$
Medium	0.592 ± 0.088	0.636 ± 0.030	0.672 ± 0.025	0.648 ± 0.057	0.582 ± 0.100	$\mathbf{0.672 \pm 0.052}$
High	0.788 ± 0.033	0.806 ± 0.040	0.814 ± 0.024	0.818 ± 0.011	0.732 ± 0.081	$\mathbf{0.836 \pm 0.036}$
Macro-F1	0.661 ± 0.062	0.715 ± 0.032	0.737 ± 0.019	0.728 ± 0.033	0.647 ± 0.053	$\mathbf{0.751 \pm 0.041}$

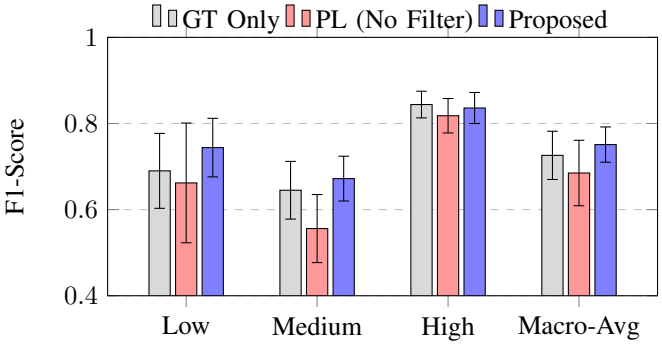

Fig. 9. Impact of Data Strategy. Comparing Ground-Truth (GT) Only against indiscriminate pseudo-labelling (No Filter) and the proposed uncertainty-aware filtering. Error bars represent standard deviation across 5 folds.

to Base (0.751) yields a consistent improvement across all metrics, confirming that both data scale and parameter count are necessary to capture the nuances of confidence.

Furthermore, the proposed architecture is compared against our preliminary work [13] which used a co-attention mechanism to fuse the acoustic feature vector with the Whisper embeddings prior to a downstream sequence model. However, as shown in Table VII, the framework proposed in this paper achieves a higher Macro-F1 score despite being architecturally simpler indicating that the previous co-attention mechanism unnecessarily increased model complexity and the proposed method achieves superior performance with greater efficiency.

V. DISCUSSION

The analysis results in Section IV suggest that the superior performance of Whisper-Base over purely acoustic self-supervised learning models such as WavLM could be due to the scale of pre-training. However, it could also be due to the nature of the training objective for these models. Unlike masked acoustic prediction, Whisper's ASR objective inherently aligns audio with linguistic semantics. This likely enables the model to learn implicit lexical information that correlates with confidence, which models trained on acoustic reconstruction will not see.

However, relying on these implicit linguistic cues creates a vulnerability when the spoken words contradict the speaker's tone. For example, a declarative statement could be misclassified as High Confidence by the baseline if it relies too heavily on the confident phrasing. The improved performance of the Hybrid model implies that the added acoustic features likely provide a corrective signal in these ambiguous instances.

To answer **RQ1**, we demonstrated that pseudo-labelling is an effective strategy for upscaling small labelled datasets. The proposed framework achieved a Macro-F1 of 0.751, outperforming the ground-truth baseline of 0.726. Moreover regarding **RQ2**, our ablation studies confirm that data quality supersedes quantity. We found that indiscriminate pseudo-labelling degraded performance compared to the baseline. This aligns with findings in [38] regarding confirmation bias where without filtering, the model reinforces its own errors on noisy samples and the Uncertainty-Aware strategy ($\tau > 0.8$) demonstrates that data quality supersedes quantity. By restricting the training set to high-confidence samples, the model was more able to learn robust decision boundaries without overfitting to the small ground truth dataset ($N = 600$).

Addressing **RQ3**, we established that fusing explicit paralinguistic features with Whisper embeddings yields a measurable performance gain over unimodal baselines. Our analysis shows that whilst Whisper provides a robust foundation, the explicit acoustic features act as a necessary corrective mechanism, al-

lowing the model to distinguish between levels of confidence. For example, the SHAP plot (Fig. 8) suggests that the model actively uses Jitter and the presence of Word repetitions to distinguish Medium from High confidence. This confirms that the feature vector acts as a corrective mechanism, preventing the semantic bias observed in the Whisper-only baseline.

Whilst the proposed architecture demonstrates strong performance in confidence estimation, several limitations exist. The primary constraint is the size of the ground truth dataset ($N = 600$). Although we mitigated this scarcity through semi-supervised learning, uncertainty filtering and pseudo-labelling, the core validation remains bound to a relatively small set of annotations on English-speaking audio meaning that the model may not generalise to speakers outside this distribution. Future work must validate these findings on larger, more diverse corpora to ensure cross-demographic fairness.

Furthermore, our model processes short, isolated audio clips between 5 and 12 seconds. Therefore, the model lacks the context needed to distinguish genuine confidence from more performative behaviours such as sarcasm. It also misses important cues from earlier in the dialogue and struggles with instances where confidence fluctuates within a single clip. Moreover, without demographic context, the model may misinterpret cultural differences in pacing or hesitation as signs of a lack of confidence rather than differences in how confidence is portrayed and received.

Finally, in this study we only focus on audio but confidence is an inherently multimodal phenomenon. Research indicates that non-verbal visual cues such as eye contact, posture, and facial micro-expressions are often stronger predictors of confidence than voice alone [55]. By relying exclusively on the audio modality, our system is blind to scenarios where the voice is steady but the body language signals doubt. Integrating visual features would likely resolve many of the ambiguities particularly observed in the Medium confidence class where acoustic cues are subtle or conflicting.

VI. Conclusion

In conclusion, this paper proposed a robust semi-supervised framework for the automatic detection of speaker confidence, addressing the dual challenges of data scarcity and label subjectivity. By introducing an uncertainty-aware pseudo-labelling mechanism, we demonstrated that it is possible to leverage unlabelled data to train reliable classifiers, provided that the training data is filtered using only high-quality data.

The analysis identifies Whisper-Base as a superior foundation compared to acoustic-only baselines like WavLM, likely due to its massive pre-training scale and implicit semantic alignment. Furthermore, our hybrid approach uses the feature vector to correct Whisper's understanding. By adding these specific details, the model can adjust its prediction, ensuring that subtle cues in the voice are not overlooked by Whisper. The ablation studies reveal that data quality supersedes quantity, and that strictly filtering pseudo-labels ($\tau > 0.8$) outperformed training on the full, noisy dataset. This confirms that for subjective tasks, training data quality is more important than indiscriminate large-scale augmentation. .

Future work will focus on the three main limitations in Section V. First, to address demographic biases, we will validate the framework on multilingual datasets and investigate how confidence markers differ across cultural groups. Moreover, we aim to integrate visual modalities (e.g., facial expressions and eye contact) to resolve the ambiguity of the Medium confidence class, we aim to move beyond analysing short, isolated clips by incorporating temporal context to measure how a speaker's confidence evolves over time.

Acknowledgments

The authors sincerely thank the seven annotators for their contribution to the dataset curation.

References

[1] R. M. Krauss and S. R. Fussell, "Social psychological models of interpersonal communication," in *Social psychology: Handbook of basic principles*, E. T. Higgins and A. W. Kruglanski, Eds. New York, NY: Guilford Press, 1996, pp. 655–701.

[2] M. Mardiana, B. Laksmana, and S. Sukardi, "Effects of self-confidence and diction on speaking skills in junior high school students," *Indo-Fintech Intellectuals: Journal of Economics and Business*, vol. 4, no. 4, pp. 1333–1344, Aug. 2024.

[3] J. J. Guyer, L. R. Fabrigar, and T. I. Vaughan-Johnston, "Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion," *Personality and Social Psychology Bulletin*, vol. 45, no. 3, pp. 389–405, 2019.

[4] A. P. Cavalcanti, A. Barbosa, R. Carvalho, F. Freitas, Y.-S. Tsai, D. Gašević, and R. F. Mello, "Automatic feedback in online learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100027, 2021.

[5] D. M. Clark and A. Wells, "A cognitive model of social phobia," in *Social phobia: Diagnosis, assessment, and treatment*, R. G. Heimberg and M. R. Liebowitz, Eds. New York: Guilford Press, 1995, pp. 69–93.

[6] X. Jiang and M. Pell, "Encoding and decoding confidence information in speech," in *Proc. Speech Prosody 2014*, 2014, pp. 573–576.

[7] H. Pon-Barry and S. M. Shieber, "Recognizing uncertainty in speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, Dec. 2010.

[8] S. Nair, M. Mohan, J. Rajesh, and P. Chandran, "On finding the best learning model for assessing confidence in speech," in *2020 The 3rd International Conference on Machine Learning and Machine Intelligence*, ser. MLMI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 58–64.

[9] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*. Springer, 2018, pp. 198–208.

[10] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP*, 2021. [Online]. Available: https://arxiv.org/pdf/2102.12394.pdf

[11] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[12] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2111.09344

[13] A. Wynn, J. Wang, and X. Tan, "Semi-supervised speech confidence detection using pseudo-labelling and whisper embeddings," in *Artificial Intelligence in Education*. Cham: Springer Nature Switzerland, 2025, pp. 266–274.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[16] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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, Apr. 2016. [Online]. Available: https://doi.org/10.1109/taffc.2015.2457417

[18] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, p. 261–292, Dec. 1996. [Online]. Available: http://dx.doi.org/10.1007/BF02686918

[19] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[20] C. Chappuis and D. Grandjean, "Set the tone: Trustworthy and dominant novel voices classification using explicit judgement and machine learning techniques," *PLOS ONE*, vol. 17, no. 6, p. e0267432, Jun. 2022. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0267432

[21] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.

[22] H. Trinh, R. Asadi, D. Edge, and T. Bickmore, "Robocop: A robotic coach for oral presentations," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 2, jun 2017.

[23] S. Chanda, K. Fitwe, G. Deshpande, B. W. Schuller, and S. Patel, "A deep audiovisual approach for human confidence classification," *Frontiers in Computer Science*, vol. 3, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.674533

[24] N. L. E. Astuti, N. N. Padmadewi, and I. N. A. J. Putra, "Speech disfluency and gestures production in undergraduate students'confidence level of speaking," *Media Bina Ilmiah*, vol. 19, no. 4, p. 4453, 2024.

[25] N. Bernstein Ratner and B. MacWhinney, "Fluency bank: A new resource for fluency research and practice," *Journal of Fluency Disorders*, vol. 56, pp. 69–80, 2018.

[26] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6089–6093.

[27] E. Boughariou, Y. Bahou, and L. H. Belguith, "Detecting speech disorders using a machine-learning guided method in spontaneous tunisian dialect speech," *SN Computer Science*, vol. 5, no. 5, Apr. 2024.

[28] P. Mohapatra, A. Pandey, B. Islam, and Q. Zhu, "Speech disfluency detection with contextual representation and data distillation," in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, ser. IASA '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 19–24.

[29] J. Liu, A. Wumaier, D. Wei, and S. Guo, "Automatic speech disfluency detection using wav2vec2.0 for different languages with variable lengths," *Applied Sciences*, vol. 13, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/13/7579

[30] H. Ameer, S. Latif, R. Latif, and S. Mukhtar, "Whisper in focus: Enhancing stuttered speech classification with encoder layer optimization," 2023.

[31] V. Tiwari, "Mfcc and its applications in speaker recognition," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:212584631

[32] M. S. Sidhu, N. A. A. Latib, and K. K. Sidhu, "MFCC in audio signal processing for voice disorder: a review," *Multimed. Tools Appl.*, 2024.

[33] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of alzheimers disease using spontaneous speech only," in *Interspeech 2021*. ISCA, Aug. 2021.

[34] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[35] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," 2022. [Online]. Available: https://arxiv.org/abs/2203.07378

[36] E. Goron, L. Asai, E. Rut, and M. Dinov, "Improving domain generalization in speech emotion recognition with whisper," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 631–11 635.

[37] M. Osman, D. Z. Kaplan, and T. Nadeem, "Ser evals: In-domain and out-of-domain benchmarking for speech emotion recognition," 2024. [Online]. Available: https://arxiv.org/abs/2408.07851

[38] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," 2020. [Online]. Available: https://arxiv.org/abs/1908.02983

[39] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," 2020. [Online]. Available: https://arxiv.org/abs/2001.07685

[40] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[41] T. Sainburg, "timsainb/noisereduce: v1.0," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[42] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. NY, USA: Association for Computing Machinery, 2010, p. 1459–1462.

[43] S. P. Bayerl, D. Wagner, E. Nöth, T. Bocklet, and K. Riedhammer, "The influence of dataset partitioning on dysfluency detection systems," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer International Publishing, 2022, pp. 423–436.

[44] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[45] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, 2018.

[46] P. Jackson and S. Haq, "Surrey Audio-Visual Expressed Emotion (SAVEE) Database," http://kahlan.eps.surrey.ac.uk/savee/Database.html.

[47] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020.

[48] Arushi, R. Dillon, and A. N. Teoh, "Real-time stress detection model and voice analysis: An integrated vr-based game for training public speaking skills," in *2021 IEEE Conference on Games (CoG)*, 2021, pp. 1–4.

[49] J. Staš, S. Ondáš, and J. Juhár, "Performance evaluation of different speech-based emotional stress level detection approaches," *IEEE Access*, vol. 13, p. 112880–112904, 2025. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2025.3584534

[50] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," 2017.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[52] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[53] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[54] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1505–1518, Oct. 2022.

[55] Y. Mori and M. D. Pell, "The look of (un)confidence: Visual markers for inferring speaker confidence in speech," *Frontiers in Communication*, vol. 4, Nov. 2019.

VII. Biography Section

Adam Wynn is a PhD student in the Department of Computer Science at Durham University. He is interested in AI in education, automatic feedback, adaptive learning and educational technology.

Jingyun Wang is an assistant professor in the Department of Computer Science at Durham University. Her research focuses on harnessing the power of AI and HCI to drive innovation in the fields of education and digital health.