Semi-Supervised Speech Confidence Detection using Pseudo-Labelling and Whisper Embeddings

Adam Wynn[1][0000−0002−1631−2151], Jingyun Wang[1][0000−0001−9325−1789], and Xiangyu Tan[2][0000−0001−9171−1788]

[1] Durham University, Durham, United Kingdom {adam.t.wynn, jingyun.wang}@durham.ac.uk
[2] Shanghai Open University, Shanghai, China tanxy@shisu.edu.cn

Abstract. Understanding speaker confidence is crucial in educational settings, as it can enhance personalised feedback and improve learning outcomes. This study introduces a novel framework for detecting speaker confidence by integrating human-engineered features with embeddings from the Whisper encoder. To address data limitations, a pseudo-labelling technique is employed to expand the labelled dataset, allowing the model to learn from both human-annotated and model-generated labels. The framework combines traditional speech features including pitch, volume, rate of speech, and the presence of disfluencies and stress, with Whisper embeddings, and uses a co-attention mechanism to fuse these representations and achieve an overall accuracy of 75%. This study contributes to advancing speech analysis, enabling applications that support personalised learning and speaking skill development.

Keywords: Confidence Detection · Disfluency Detection · Semi-Supervised Learning · Speaking Skills

1 Introduction

Effective communication is essential in education, shaping how knowledge is shared and understood [12]. Confidence, in particular, influences a student's clarity, credibility, and engagement when speaking [20], making it vital for tasks like presentations and public speaking. Understanding confidence can support automatic feedback and help educators identify areas where support is needed [4] to help improve communication skills.

Prior research has demonstrated that confidence is reflected in acoustic features such as pitch, speech rate, and vocal intensity [10]. However, earlier work relied on manual annotation and small-scale studies [31], and despite advances in AI [22], confidence detection remains underexplored due to a lack of annotated datasets. Therefore, this study explores the detection of speaker confidence, by proposing a novel semi-supervised framework that integrates human-engineered features including pitch and speech rate, with embeddings from the Whisper encoder [26] using a co-attention mechanism. Human-engineered features are used to generate pseudo-labels [16] for unlabelled data to expand the dataset and improve the generalisability of the model.

This work aims to answer the following research questions: **RQ1:** How can semi-supervised learning and model-based pseudo-labelling be leveraged to address the scarcity of confidence-labelled data? **RQ2:** To what extent do pitch, rate of speech, amplitude and the presence of disfluencies and stress, influence the model's perception of confidence? Our main contribution is introducing a semi-supervised framework to detect speaker confidence, offering a solution for assessing students' verbal communication skills and enabling adaptive, real-time feedback in educational settings.

2 Related Work

Confidence in speech is generally perceived through a combination of vocal characteristics including pitch, volume, speech rate, and clarity. These features act as the foundation for evaluating speaker confidence in both human and automated systems. Prior research [11] discovers that confident expressions have highest f0 range, mean amplitude and amplitude range and unconfident expressions are highest in mean f0, slowest in speaking rate, with more frequent pauses. Automated systems designed for confidence detection have been proposed including providing automated feedback to users rehearsing oral presentations based on speech quality, content coverage and audience reaction [30], and predicting the confidence of a speaker using Mel-Frequency Cepstral Coefficients (MFCC) as inputs based on clarity, modulation, pace, and volume.

Confidence can also influence the frequency of speech disfluencies and a lack of confidence often results in more frequent pauses and fillers [2]. Several approaches for disfluency classification have recently emerged [14][3], but require large labelled datasets. To address this challenge, Mohapatra et al. [21] 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. Moreover, Ameer et al. [1] used the Whisper model for multi-class disfluency classification, introducing an encoder-freezing strategy, outperforming Wav2Vec 2.0 models.

Confidence detection is also closely linked to Speech Emotion Recognition (SER), as both fields rely on analysing the speakers' prosody. Pepino et al. [24] proposed using Wav2vec 2.0 embeddings for SER by combining the output of several layers of pre-trained Wav2Vec 2.0 using trainable weights to produce richer speech representations. Additionally, they integrated prosodic features into the model, including pitch and loudness, resulting in further improvements in performance. Goel et al [6] proposed to improve the generalisability of SER models by introducing CAMuLENET and incorporating attention which outperforms baseline Whisper, Wav2Vec 2.0 and HuBERT and improves the generalisation to unseen speakers. These advancements highlight the potential of deep learning in enhancing SER and confidence detection by effectively extracting prosodic cues embedded in speech to provide more accurate and generalisable predictions.

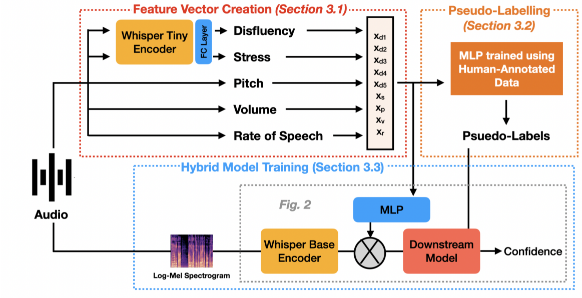

Fig. 1. Overall Pipeline of the Confidence Classification System during Training

3 Methodology

Building on the advancements of using deep learning for feature extraction and the classification of audio, this study proposes a novel framework for classifying speaker confidence which involves a pipeline (Fig. 1) that integrates feature vectors with audio embeddings using a pre-trained Whisper-base model encoder. Both representations are fused through a co-attention mechanism to incorporate information from both sources effectively and develop a robust system for classifying speaker confidence levels on limited annotated data.

Due to the lack of publicly available data labelled for confidence, a manual annotation system was developed to classify audio clips into low, medium, and high confidence. Three fluent English speakers (a native male, a non-native female, and a non-native female English speech expert) rated each clip, with final labels based on averaged scores. The resulting dataset includes 444 clips (247 male, 197 female), drawn from TEDLIUM [8], SEP-28K [15], and non-native English assessments, comprising 172 high-, 151 medium-, and 121 low-confidence samples. This dataset serves as the human-annotated test set for evaluating performance on unseen data. To compensate for its limited size, a model-based pseudo-labelling technique is used to expand the dataset for training and validation. Further details are provided in Section 3.2.

3.1 Feature Vector Creation

A 9-dimensional feature vector was engineered for each audio clip to capture prosodic cues linked to speaker confidence, including pitch and amplitude variation, rate of speech, presence of stress, and five types of speech disfluencies: word repetitions, prolongations, interjections, blocks, and sound repetitions. Prior studies link pitch, amplitude, and speaking rate to perceived confidence [11][7], whilst disfluencies and stress have been associated with low confidence [2][23].

In this study, pitch variation is measured using the SPICE pitch tracker [5], and amplitude is calculated using the normalised variation of the amplitude envelope, which reflects the intensity and energy of speech delivery. The rate of speech is derived using the MyProsody [28] library.

The Disfluency and stress detection models share a common model architecture. In order to identify disfluencies, data from the SEP-28K [15] and FluencyBank [27] datasets is used. Only the audio clips where all three raters agreed were used and data was excluded which had poor audio quality. The dataset was augmented and balanced by applying pitch shifting and Gaussian noise to the disfluent audio files. The model architecture consists of the Whisper Tiny Encoder [26], chosen due to its ability to generalise compared to the base model, which appeared to overfit the data.

For the stress detection task, a possible solution to the lack of sufficient labelled data is to establish a mapping between emotional states and stress levels. Research indicates that the emotions sadness, fear, and anger are most closely associated with high stress levels [29]. To address this, audio segments were extracted from the RAVDESS [18], SAVEE [9] and TESS [25] SER datasets, where the labels for sadness, fear, and anger were relabelled as "stress", while the remaining labels were labelled as "neutral".

3.2 Model-Based Pseudo-Labelling

To address the limitation of a small ground truth dataset, a model-based pseudo-labelling approach was used to generate additional labelled data. After the feature vectors 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. This process did not involve the audio or the Whisper embeddings directly in an attempt to prevent data leakage.

This model was initially trained on 363 samples with Adam optimiser [13] and cross-entropy loss, and achieved an accuracy of 79.19% using 10 fold validation. The trained MLP was then applied to a larger, unlabelled dataset comprising 2640 audio samples (880 clips per label after downsampling to ensure equal class sizes) to generate pseudo-labels. 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.

3.3 Hybrid Model Training

The hybrid model (Fig. 2) integrates information from both the engineered feature vector and audio embeddings derived from Whisper-base. The training process uses the pseudo-labelled dataset for training and validation, whilst the original ground truth dataset was reserved for testing.

First, the engineered feature vector was input into a MLP different from the one used during pseudo-labelling, to generate a 128-dimensional embedding. Simultaneously, the pre-processed audio data was input into the Whisper-base encoder to extract high-dimensional embeddings from the last encoder layer. These two representations were fused using a co-attention mechanism, where attention weights were applied to the Whisper embedding. The combined embeddings were passed through a downstream network to predict confidence levels.

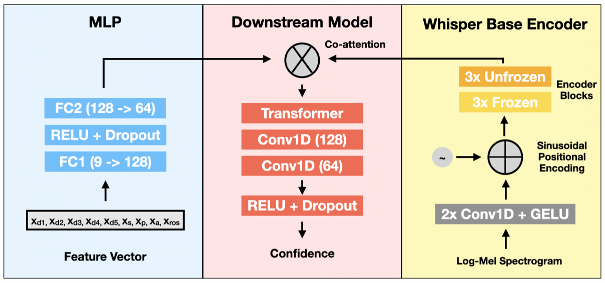

Fig. 2. Confidence Detection using Whisper-Base Encoder and Feature Vector

4 Results and Discussion

4.1 Disfluency and Stress Detection

The disfluency classification model (Fig. 1) was trained for 50 epochs using AdamW Optimiser [19] with a learning rate of 2.5×10^{-5} and binary cross-entropy loss, using 10-fold cross-validation and early stopping. The model performed best on interjections (Acc: 0.80, F1: 0.80) and prolongations (Acc: 0.78, F1: 0.77), followed by sound repetitions (Acc: 0.74, F1: 0.74), word repetitions (Acc: 0.68, F1: 0.68) and blocks (Acc: 0.69, F1: 0.68). These findings align with existing literature, indicating that detecting blocks and word repetitions poses challenges [21] whilst interjections are more easily identified [17]. On the other hand, the stress classification model was trained with the same hyperparameters as above and achieved an accuracy of 0.86 (F1: 0.85).

4.2 Confidence Detection

The Confidence classification model, as shown in Fig. 2, was trained for 200 epochs using the AdamW optimiser [19], with an initial learning rate of $2.5{*}10e^{-5}$ using cross entropy loss. Training and validation used data from the pseudo-labelled data and the model was tested on the human-annotated data only.

	Accuracy	F1	Precision	Recall
Low Confidence	0.88	0.80	0.73	0.88
Medium Confidence	0.61	0.67	0.74	0.62
High Confidence	0.78	0.79	0.79	0.78
Overall	0.75	0.75	0.75	0.75

Table 1. Results for the Confidence Classification Model

As shown in Table 1, the model achieves stronger results for low (Acc: 0.88, F1: 0.80) and high (Acc: 0.78, F1: 0.79) confidence, but struggles with medium confidence (Acc: 0.61, F1: 0.67), perhaps because medium confidence is more ambiguous and can be misclassified as either low or high confidence. After training the model, SHapley Additive exPlanations (SHAP) identified Pitch Variation, Amplitude Variation, and Sound Repetitions as the most important features. As shown in Fig. 3, sound repetitions strongly increased the likelihood of classifying the speaker as low confidence, though the effect on high confidence was less clear. For pitch, higher variation typically correlated with negative SHAP values for both low and high confidence, suggesting that the model associates high pitch variation with medium confidence. Amplitude variation, the most influential feature, was less clear, but in general, samples with lower amplitude variation were less likely to indicate low confidence.

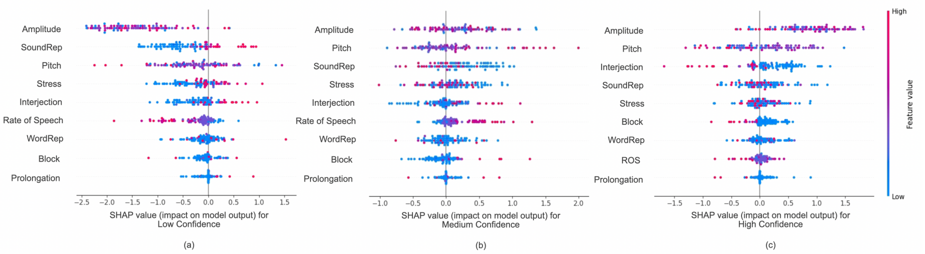

Fig. 3. Shap Values for Low, Medium and High Confidence Respectively

5 Conclusion

This study presents a novel framework for detecting speaker confidence in audio, combining feature engineering and pseudo-labelling to create a robust classification system. By training a model to generate pseudo-labels, the final hybrid model can incorporate these pseudo-labels during training to differentiate between confidence levels enabling the model to generalise its learning, resulting in stronger performance when applied to human-annotated data.

A key limitation is that the pseudo-labelling process depends on the quality of human-annotated data. Biases in this data can introduce label noise and although the hybrid model doesn't directly use the annotations, completely unseen data would support more robust training. Enhancing the dataset with more annotators and diverse sources would also improve generalisability. Another limitation is the lack of consideration for linguistic and cultural differences in confidence expression. Further testing across languages and cultures will be necessary to ensure the model can generalise effectively across cultures.

In the future, this framework could support educational applications, such as assessing student confidence in oral presentations and providing targeted feedback. Moreover, the framework could be used to predict other abstract communication skills beyond confidence, such as persuasiveness or empathy. This study lays the groundwork for advancing speech analysis systems and paves the way for applications that allow for personalised speaking skills development.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ameer, H., Latif, S., Latif, R., Mukhtar, S.: Whisper in focus: Enhancing stuttered speech classification with encoder layer optimization (2023)
2. Astuti, N.L.E., Padmadewi, N.N., Putra, I.N.A.J.: Speech disfluency and gestures production in undergraduate students'confidence level of speaking. Media Bina Ilmiah **19**(4), 4453–4462 (2024)
3. Bougariou, E., Bahou, Y., Belguith, L.H.: Detecting speech disorders using a machine-learning guided method in spontaneous tunisian dialect speech. SN Computer Science **5**(5) (Apr 2024). https://doi.org/10.1007/s42979-024-02775-8
4. Cavalcanti, A.P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.S., Gašević, D., Mello, R.F.: Automatic feedback in online learning environments: A systematic literature review. Computers and Education: Artificial Intelligence **2**, 100027 (2021). https://doi.org/10.1016/j.caeai.2021.100027
5. Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., Velimirović, M.: Spice: Self-supervised pitch estimation. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **28**, 1118–1128 (apr 2020). https://doi.org/10.1109/TASLP.2020.2982285
6. Goel, A., Hira, M., Gupta, A.: Exploring multilingual unseen speaker emotion recognition: Leveraging co-attention cues in multitask learning (2024), https://arxiv.org/abs/2406.08931
7. Guyer, J.J., Fabrigar, L.R., Vaughan-Johnston, T.I.: Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. Personality and Social Psychology Bulletin **45**(3), 389–405 (2019)
8. Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., Esteve, Y.: 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. pp. 198–208. Springer (2018)
9. Jackson, P., Haq, S.: Surrey Audio-Visual Expressed Emotion (SAVEE) Database — kahlan.eps.surrey.ac.uk. http://kahlan.eps.surrey.ac.uk/savee/Database.html, [Accessed 17-02-2025]
10. Jiang, X., Pell, M.: Encoding and decoding confidence information in speech. In: Proc. Speech Prosody 2014. pp. 573–576 (2014). https://doi.org/10.21437/SpeechProsody.2014-103
11. Jiang, X., Pell, M.D.: The sound of confidence and doubt. Speech Communication **88**, 106–126 (2017). https://doi.org/https://doi.org/10.1016/j.specom.2017.01.011
12. Kasemsap, K.: Digital storytelling and digital literacy. In: Advances in Educational Marketing, Administration, and Leadership, pp. 151–171. IGI Global (2017)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), https://arxiv.org/abs/1412.6980
14. Kourkounakis, T., Hajavi, A., Etemad, A.: Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6089–6093 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053893
15. Lea, C., Mitra, V., Joshi, A., Kajarekar, S., Bigham, J.: Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In: ICASSP (2021), https://arxiv.org/pdf/2102.12394.pdf
16. Lee, D.H.: 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. Liu, J., Wumaier, A., Wei, D., Guo, S.: Automatic speech disfluency detection using wav2vec2.0 for different languages with variable lengths. Applied Sciences **13**(13) (2023). https://doi.org/10.3390/app13137579, https://www.mdpi.com/2076-3417/13/13/7579
18. Livingstone, S.R., Russo, F.A.: 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 **13**(5) (2018). https://doi.org/10.1371/journal.pone.0196391
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
20. Mardiana, M., Laksmana, B., Sukardi, S.: Effects of self-confidence and diction on speaking skills in junior high school students. Indo-Fintech Intellectuals: Journal of Economics and Business **4**(4), 1333–1344 (Aug 2024)
21. Mohapatra, P., Pandey, A., Islam, B., Zhu, Q.: Speech disfluency detection with contextual representation and data distillation. In: Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications. p. 19–24. IASA '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3539490.3539601
22. Nair, S., Mohan, M., Rajesh, J., Chandran, P.: On finding the best learning model for assessing confidence in speech. In: 2020 The 3rd International Conference on Machine Learning and Machine Intelligence. p. 58–64. MLMI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3426826.3426838, https://doi.org/10.1145/3426826.3426838
23. Ningrum, N., Listyani, L.: Academic speaking students' efforts in minimizing their lack of self- confidence. Prominent **5**, 141–167 (07 2022). https://doi.org/10.24176/pro.v5i2.7874
24. Pepino, L., Riera, P., Ferrer, L.: Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In: Proc. Interspeech 2021. pp. 3400–3404 (2021). https://doi.org/10.21437/Interspeech.2021-703
25. Pichora-Fuller, M.K., Dupuis, K.: Toronto emotional speech set (TESS) (2020)
26. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022), https://arxiv.org/abs/2212.04356
27. Ratner, N.B., MacWhinney, B.: Fluency bank: A new resource for fluency research and practice. Journal of fluency disorders **56**, 69–80 (2018)
28. Shahabks: Shahabks/myprosody: A python library for measuring the acoustic features of speech (simultaneous speech, high entropy) compared to ones of native speech. https://github.com/Shahabks/myprosody (2021), accessed: 03 June 2024
29. Staš, J., Hládek, D., Sokolová, Z., Čech, M., Škotková, K., Poremba, P.: Analysis and detection of speech under emotional stress. In: 2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA). pp. 493–498. IEEE (2023)
30. Trinh, H., Asadi, R., Edge, D., Bickmore, T.: Robocop: A robotic coach for oral presentations. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(2) (jun 2017). https://doi.org/10.1145/3090092, https://doi.org/10.1145/3090092
31. Williams, G., McLellan, B., Sivesind, G.: Identifying Confidence in Speech p. 6 (2017)