Exploring Sentence Stress Detection Leveraging Whisper-based Speech Foundation Models

Ting-An Hung[*1], **Yu-Hsuan Hsieh**[*1], **Tien-Hong Lo**[1], **Yung-Chang Hsu**[2], **Berlin Chen**[1]

[1]**National Taiwan Normal University,** [2]**EZAI Corp.**
{annhung323451, ivyhsieh0118}@gmail.com,
{teinhonglo, berlin}@ntnu.edu.tw, mic@ez-ai.com.tw

Abstract

Sentence stress reflects the relative prominence of words within a sentence. It is fundamental to speech intelligibility and naturalness, and is particularly important in second language (L2) learning. Accurate stress production facilitates effective communication and reduces misinterpretation. In this work, we investigate Sentence Stress Detection (SSD) using Whisper-based Transformer speech models under diverse settings, including model scaling, backbone–decoder interactions, architectural and regularization enhancements, and embedding visualization for interpretability. Results show that smaller Whisper variants outperform larger ones under limited data. With architectural and regularization enhancements, and by fixing a decoder whose capacity matches the dataset scale, both small and large backbones benefit. Consequently, even larger models can achieve competitive or superior performance under data-scarce conditions, partially mitigating data limitation effects. Embedding analysis reveals clear separation between stressed and unstressed words. These findings offer practical insights into model selection, architecture design, and interpretability for SSD applications, with implications for L2 learning support tools.

1 Introduction

Automatic detection of sentence stress in spoken language is crucial for speech intelligibility, prosodic naturalness, and perceived fluency, particularly in second language (L2) learning (Ladd, 2008; Lee et al., 2016; van Heuven, 2014). Misplaced stress in L2 learners can lead to misunderstandings and reduced comprehension, motivating the development of automated Sentence Stress Detection (SSD) systems for assessment and feedback (Lin et al., 2020; Kakouros and Räsänen, 2016).

Recent advances in pre-trained speech foundation models, such as Whisper, enable the extraction of rich embeddings that encode both acoustic and prosodic information (Radford et al., 2022; Bain et al., 2023). Whisper models, trained on massive multilingual corpora, can be adapted to downstream tasks like SSD without requiring extensive task-specific data (Nguyen et al., 2023; de Seyssel et al., 2023). Building on this, the WhiStress model (Yosha et al., 2025) demonstrated the effectiveness of Whisper embeddings for prosodic feature learning. However, systematic studies investigating how model size, architectural choices, and regularization strategies affect SSD performance and interpretability remain limited.

To this end, we explore Whisper-based SSD under diverse settings, including model scaling, backbone–decoder interactions, and architectural enhancements. We also analyze embedding representations for interpretability (Van Heuven, 2018; Arvaniti, 2020).
Our main contributions are as follows:

- evaluating Whisper-based SSD across multiple model sizes,
- analyzing the impact of decoder configuration, architectural enhancements, and regularization,
- conducting embedding visualization to interpret stress representations.

2 Related Work

2.1 Sentence Stress Detection

Early studies on SSD relied on handcrafted acoustic-prosodic features such as pitch (F0), intensity, and duration, modeled using Support Vec-

[*]Equal contribution.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 314–319

tor Machines (SVMs) or Hidden Markov Models (HMMs).(Mishra et al., 2012; Auran et al., 2004). While effective in constrained settings, these methods required extensive domain expertise and failed to capture complex interactions among prosodic cues. Subsequent deep learning approaches, using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformer architectures (Vaswani et al., 2017), advanced SSD by learning hierarchical features directly from raw speech (Baevski et al., 2020; Pasad et al., 2021). However, these models remain data-intensive and often lack interpretability, particularly in low-resource L2 contexts.

2.2 Pre-trained Speech Models and Whisper

Self-supervised models such as Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) learn general-purpose speech representations that encode both phonetic and prosodic cues, facilitating transfer to downstream tasks. More recently, Whisper (Radford et al., 2022) introduced a large-scale encoder–decoder architecture trained on massive multilingual corpora. Whisper embeddings have demonstrated utility beyond ASR, including emotion recognition and prosodic analysis (Nguyen et al., 2023). Building on this foundation, the WhiStress model (Yosha et al., 2025) adapted Whisper for SSD using an alignment-free framework, but its evaluation was limited to a single variant (Whisper-small.en). Broader studies examining scaling behavior, architectural design, and regularization effects remain scarce.

Although pre-trained speech models have significantly advanced prosodic modeling, there has been limited investigation into how model size, architecture, and regularization influence SSD performance and interpretability.

3 Method

3.1 Model Architectures and Configurations

Our architecture, which is similar to WhiStress (Yosha et al., 2025), is based on the Whisper encoder to extract speech embeddings that implicitly encode acoustic and prosodic cues. A stress decoder then predicts word-level stress labels. The overall architecture is shown in Figure 1. Compared with WhiStress, which directly applies Whisper embeddings to a classification head, our design introduces several modifications: Backbone–decoder scaling:

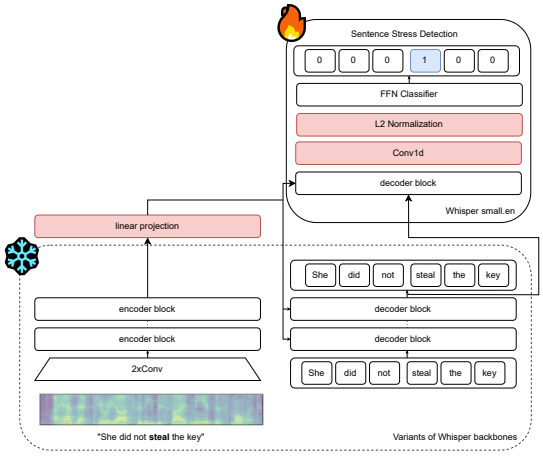

Figure 1: Overall architecture of the proposed SSD model. Input speech is processed by the Whisper encoder, followed by a fixed or trainable stress decoder. Optional components, including Conv1D, L2 normalization, and dropout, are applied depending on the model configuration.

varying both encoder size and decoder capacity. Following WhiStress (Yosha et al., 2025), which utilized the 9th layer out of 12 encoder layers in Whisper-small for optimal stress representation, intermediate-to-upper encoder layers were found to capture prosodic and phonetic cues most effectively. Accordingly, We generalize this approach by selecting approximately three-quarters of the encoder layers for each Whisper variant, namely 3, 5, 9, 18, and 24 layers for the tiny, base, small, medium, and large models, respectively.

This proportional selection strategy aims to preserve high-level prosodic features while maintaining training efficiency and mitigating overfitting risks.

We incorporate architectural enhancements by adding Conv1D and projection layers to better capture local prosodic dynamics. We also apply regularization mechanisms, including dropout and L2 normalization, to improve training stability and model generalization.

3.2 Experimental Configurations

Configuration I: Joint Scaling of Backbone and Stress Decoder. In this configuration, both the Whisper backbone and the stress decoder vary between Base, Small, Medium, and Large (Radford et al., 2022). This allows us to examine how the overall model capacity affects SSD, including po-

tential overfitting for larger models.

Configuration II: Fixed Decoder, Varying Backbone. Here, the stress decoder is fixed as Whisper-Small.en (Radford et al., 2022), while the backbone varies in size. This isolates the contribution of the backbone to SSD performance, providing a fair comparison of different representation capacities without confounding changes in the classification head.

Configuration III: Architectural and Regularization Enhancements. The backbone is fixed at Whisper-Small.en. The stress decoder incorporates several enhancements:

- **Conv1D layer**: captures local temporal dependencies in frame-level embeddings, enhancing local prosodic pattern learning.
- **L2 normalization**: word-level embeddings x are normalized as $\hat{x} = x/\|x\|_2$, standardizing embedding magnitudes to improve generalization and stability.
- **Dropout**: randomly zeros out portions of embeddings during training to prevent overfitting, especially important for high-dimensional embeddings.

Configuration IV: Embedding Visualization and POS Analysis. We use t-SNE to project word-level embeddings and inspect stress clustering. Part-of-speech (POS) analysis examines whether specific word types, such as nouns, verbs, or function words, are more challenging, providing insight into systematic error patterns.

4 Experiments

4.1 Dataset

All experiments are conducted on the TINYSTRESS-15K dataset (Eldan and Li, 2023), a fully synthetic English speech corpus designed for SSD evaluation. It contains 15,000 training samples and 1,000 test samples, totaling approximately 15 hours of audio. Word-level stress annotations and precise time alignment are provided. Prosodic parameters such as pitch, duration, and amplitude are manipulated to simulate natural sentence stress, while multiple synthetic speaker voices increase variability. This controlled synthetic design allows for consistent evaluation of model performance under diverse prosodic variations without the need for costly manual labeling. However, as the dataset is fully synthetic, it may not perfectly capture the acoustic nuances of natural speech. Future work will include testing on natural speech corpora to further validate model performance.

4.2 Training Details

All experiments are trained for 20 epochs with batch size 16, using the AdamW optimizer (Kingma and Ba, 2015) with an initial learning rate of $1e{-}4$ and cross-entropy loss. Whisper backbones are frozen, and only the stress decoder is updated. The training set is split into 90% for training and 10% for validation. No early stopping is applied; the model achieving the best F1 score on the validation set among all 20 epochs is used for reporting results. Models are evaluated using word-level F1 score, precision, and recall.

4.3 Results

Table 1 summarizes the performance of different backbone-decoder combinations. To isolate the effect of backbone representation power, we fix the decoder as Whisper-Small.en and vary the backbone size, as shown in Table 2.

Model	Precision	Recall	F1
Tiny	0.8733	0.8576	0.8653
Base	0.8834	0.8885	0.8859
Small	**0.9301**	**0.9288**	**0.9294**
Medium	0.8399	0.8381	0.8390
Large	0.7309	0.8245	0.7748

Table 1: Configuration I: Joint scaling of backbone and stress decoder. Larger models do not necessarily improve performance, likely due to overfitting.

Backbone	Precision	Recall	F1
Tiny	0.8664	0.8957	0.8808
Base	0.8726	0.9065	0.8892
Small	0.9348	0.9187	0.9267
Medium	**0.9509**	**0.9612**	**0.9560**

Table 2: Configuration II: Fixed decoder (Small.en) with varying backbone sizes. Small backbone provides optimal trade-off between capacity and generalization, while Medium achieves the highest F1.

5 Discussion

Configuration I: Joint Scaling Larger models do not consistently improve SSD performance; the Large backbone shows overfitting under limited data. Small and Base achieve better generalization, suggesting that model capacity must be

Enhancement	Precision	Recall	F1
Baseline	0.8664	0.8957	0.8808
Dropout only	0.9095	0.9324	0.9208
L2 normalization only	0.9125	0.8928	0.9025
Conv1D only	0.9209	0.9302	0.9256
Conv1D + L2 normalization	0.9366	0.9144	0.9254
Dropout + L2 normalization	0.8941	0.9295	0.9115
Dropout + Conv1D	**0.9364**	**0.9317**	**0.9340**

Table 3: Configuration III: Ablation study of architectural and regularization enhancements on SSD performance using the Whisper-Tiny backbone with a fixed Small decoder.

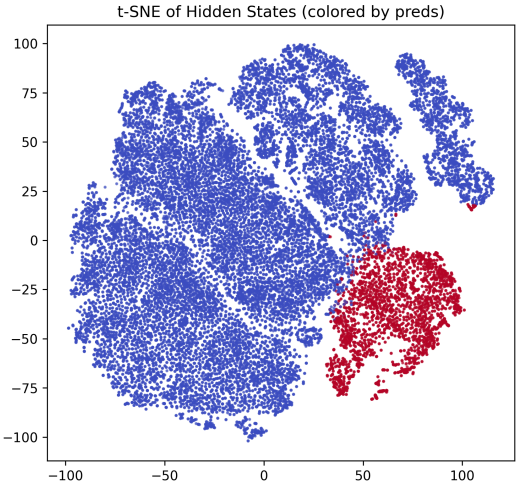

Figure 2: Configuration IV: t-SNE visualization of word-level embeddings. Red points = stressed words, blue = unstressed. POS analysis indicates nouns and verbs cluster more distinctly than function words, suggesting systematic differences in classification difficulty.

matched with data scale. The lack of linear projection may further limit larger models, as seen in the improvements from architectural enhancements in Configuration III.

Configuration II: Fixed Decoder, Varying Backbone With the decoder fixed, the medium backbone achieves the best F1 (0.9560), confirming that the backbone size directly impacts SSD quality. Small still balances accuracy and efficiency, making it practical in resource-constrained settings.

Configuration III: Architectural Enhancements Ablation in the tiny backbone shows that Conv1D (F1 = 0.9256), Dropout (0.9208) and L2 normalization (0.9025) each improve performance over baseline (0.8808). Additional analysesexamining partial module combinations reveal that: using only Dropout + L2 normalization yields F1 = 0.9115; only Conv1D + L2 normalization gives F1 = 0.9253; and only Conv1D + Dropout results in F1 = 0.9340, making it the bbest performingpartial module combination. Gains on larger backbones were marginal, indicating current performance may be bounded by dataset size. These targeted enhancements remain crucial for stable training on limited data.

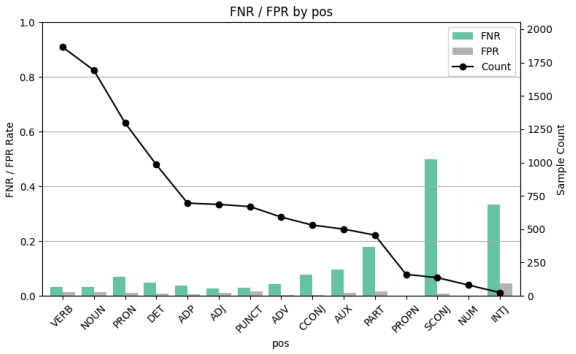

Figure 3: FNR/FPR by POS with sample counts (line). High-frequency POS (VERB/NOUN) show low FNR and near-zero FPR, while low-frequency categories such as SCONJ and NUM exhibit higher FNR despite smaller counts.

Configuration IV: Embedding and POS Analysis t-SNE visualizations of word-level embeddings show a clear separation between stressed and unstressed words, with stressed words forming more compact clusters. POS analysis further reveals that function words are more error-prone compared to content words. In a more detailed breakdown, high-frequency categories such as VERB, NOUN, PRON, DET, and ADP exhibit low FNR and near-zero FPR, indicating reliable predictions, while low-frequency categories like SCONJ, NUM, and INTJ have high FNR but low FPR, meaning many true instances are missed but mislabeling is rare. These observations suggest that improving VERB prediction and increasing data for rare POS, as well as incorporating POS-aware modeling or additional prosodic cues, could enhance the overall performance.

6 Conclusion and Future Work

This study systematically investigated Sentence Stress Detection (SSD) using Whisper-based models, focusing on model scaling, decoder configuration, architectural enhancements, and embed-

ding interpretability. Results show that scaling backbone and decoder simultaneously may cause overfitting under limited data, while fixing the decoder provides a clearer evaluation of backbone capacity, benefiting both small and large backbones. Consequently, larger models can achieve competitive or superior performance under data-scarce conditions, partially mitigating data limitation effects. Lightweight modifications such as Conv1D, L2 normalization, and dropout improve robustness, and embedding analyses reveal both stress separability and systematic misclassification patterns, particularly for smaller backbones. POS analysis indicates that function words are more challenging, suggesting potential benefits from POS-aware modeling or additional prosodic cues.

Future work will extend SSD to multilingual and cross-lingual contexts, incorporate richer linguistic features (e.g., syllable structure, phonological rules, POS embeddings), and evaluate real-world scenarios including noisy, spontaneous, and accented speech. Multi-layer embedding fusion may further capture complementary prosodic cues. Finally, given current performance appears constrained by dataset scale, exploring data-efficient strategies such as semi-supervised learning, augmentation, or active learning will be critical to overcoming data scarcity and advancing SSD performance.

Overall, this work provides practical insights for designing SSD models and informs the development of L2 learning support tools, offering both quantitative and qualitative guidance for future research.

References

Amalia Arvaniti. 2020. The phonetics of prosody.

Cyril Auran, Caroline Bouzon, and Daniel J. Hirst. 2004. The aix-marsec project: an evolutive database of spoken british english. *Speech Prosody 2004*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Vincent J. van Heuven. 2014. Acoustic correlates and perceptual cues of word and sentence stress: Mainly english and dutch. In *INTERSPEECH 2014*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sofoklis Kakouros and Okko Räsänen. 2016. 3pro – an unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, 82:67–84.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

D. Robert Ladd. 2008. *Intonational Phonology*, 2 edition. Cambridge Studies in Linguistics. Cambridge University Press.

Gary Lee, Ho-Young Lee, Jieun Song, Byeongchang Kim, Sechun Kang, Jinsik Lee, and Hyosung Hwang. 2016. Automatic sentence stress feedback for non-native english learners. *Computer Speech & Language*, 41.

Binghuai Lin, Liyuan Wang, Xiaoli Feng, and Jinsong Zhang. 2020. Joint detection of sentence stress and phrase boundary for prosody. In *INTERSPEECH*, pages 4392–4396.

Taniya Mishra, Vivek Rangarajan Sridhar, and Alistair Conkie. 2012. Word prominence detection using robust yet simple prosodic features. In *Interspeech 2012*, pages 1864–1867.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, and Michael Hassid. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2023. Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *arXiv preprint arXiv:2312.14069*.

Vincent Van Heuven. 2018. *Acoustic Correlates and Perceptual Cues of Word and Sentence Stress: Theories, Methods and Data*, pages 15–59.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Iddo Yosha, Dorin Shteyman, and Yossi Adi. 2025. Whistress: Enriching transcriptions with sentence stress detection. In *Interspeech 2025*.