Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SupraDoRAL: Automatic Word Prominence Detection Using Suprasegmental Dependencies of Representations with Acoustic and Linguistic Context

Jhansi Mallela[1], *Upendra Vishwanath Y.S.*[2], *Sankara Bharadwaj Rangavajjala*[1], *Bhaskar Bhatt*[1], *Chiranjeevi Yarra*[1]

[1]Speech Processing Lab, Language Technologies Research Center, IIIT Hyderabad, India
[2]Hewlett Packard Enterprise, Bangalore, India

jhansi.mallela@research.iiit.ac.in, upendravishwanath7@gmail.com, chiranjeevi.yarra@iiit.ac.in

Abstract

Word-level prominence plays a crucial role in spoken language understanding, as it helps to interpret the speaker's intent. However, automatic word-level prominence detection remains underexplored, particularly in non-native speech, where prominence patterns are more variable due to native language influence. While prominence is determined by acoustic features such as energy, duration, and pitch, capturing statistics on these acoustics at the word level provides only a global representation, missing finer suprasegmental variations. In this study, we consider syllables as the suprasegmental unit and propose a methodology that jointly models syllable prominence variations and their contribution to word prominence using sequential neural networks. This enables learning word-level prominence representations with minimal labeled data. Our method outperforms an unsupervised n-gram-based baseline by 24.01% and a supervised SVM by 5.73%, demonstrating its effectiveness over both approaches.

Index Terms: prominence detection, spoken language understanding, suprasegmental dependencies, non-native

1. Introduction

Word prominence in speech refers to the emphasis placed on a word within a sentence [1, 2, 3]. It helps speakers highlight key message parts and indicate intent, influencing meaning even when the words remain the same [4]. Listeners rely on prosodic cues beyond literal meaning, making prominence crucial in spoken language understanding (SLU) [5]. Automatic prominence detection benefits applications such as voice assistants for natural interactions, expressive text-to-speech (TTS) for improved prosody, spoken language summarization for highlighting key information [6, 7]. In Computer-Assisted Language Learning (CALL) systems [8], prominence detection is essential for non-native English learners, who often struggle with prosodic emphasis due to native language influence [9]. Misplaced emphasis can lead to confusion between syllable-level and word-level prominence, affecting intelligibility and meaning. CALL systems that provide word prominence feedback can help learners improve fluency and effective communication.

Prominence is reflected through acoustic parameters such as energy, duration, and f0 variations [4]. The prominent words have higher energy and duration compared to the non-prominent words in a sentence. Prominence detection in speech has been widely studied using various approaches, primarily relying on acoustic and lexical features. Traditional methods often aggregate word-level acoustic descriptors, such as fundamental frequency (F0), energy, duration, and spectral tilt, to determine prominence [4, 10, 11]. In [12], authors employed statistical models that used handcrafted acoustic and lexical features.

Modeling approaches in literature: In terms of modeling, both supervised and unsupervised approached are explored. Unsupervised methods estimated prominence by evaluating the unpredictability of speech using language models (LMs) and acoustic models (AMs) at the word level [4, 13]. In these models, words with lower predictability were considered more prominent. Supervised methods, such as support vector machines (SVMs) and k-nearest neighbors (KNNs), incorporated language model probabilities and part-of-speech tags along with aggregated acoustic descriptors [14]. However, these approaches operated at the word level, losing contextual and prosodic dependencies across syllables. In [14], unsupervised n-gram models were explored to capture word prominence by analyzing frame-level sequential patterns. These models represented prosodic features such as F0, energy, and duration using cluster indices replacing continuous variations. Prominence was inferred based on the predictability of these sequences, with higher unpredictability indicating prominent words. Recurrent neural networks (RNNs), particularly bidirectional RNNs (BRNNs), were later introduced to model temporal dependencies across words in prominence detection [15]. However, BRNNs primarily rely on word-level aggregated features, limiting their ability to capture fine-grained suprasegmental variations within words.

Motivation and limitations in literature: Prominence perception in speech is determined by the emphasis placed on syllables and words. In English, each word (content words) has one prominent syllable [16], but this does not always make the entire word prominent in a sentence. Word prominence depends on syllabic stress patterns and the word's contextual role [1, 2]. While native speakers intuitively distinguish between syllable and word prominence, non-native speakers often struggle due to differences in their native rhythmic structures. This can lead to misplaced stress, where entire words are emphasized instead of specific syllables, making prominence detection more challenging. Such misplacement further complicates spoken language understanding (SLU) systems, which rely on accurate stress patterns for interpretation.

Existing approaches to prominence detection primarily focus on word-level aggregates or dependencies between words in a sentence [15]. While these methods are effective for native speakers, they overlook syllabic stress patterns that are crucial in non-native speech, where incorrect syllable emphasis alters word prominence perception. Although n-gram modeling captures frame-level dependencies [14], it does not capture dependencies among syllables, which are the smallest units in speech where prominence is perceived. However, frames are too short to carry perceptual information for prominence detection. As a result, detecting word prominence in non-native speech re-

10.21437/Interspeech.2025-2519

quires a finer-grained approach that accounts for syllable stress dependencies, rather than relying solely on word-level features or frame-level modeling.

Beyond the choice of modeling approaches, many studies on prominence detection rely on handcrafted acoustic features, while some incorporate POS tags and syntactic dependencies [14]. However, prominence perception depends on both acoustic and linguistic context [17], which these approaches fail to fully capture. Self-supervised models like wav2vec 2.0 and HuBERT encode both acoustic and linguistic information from large-scale speech data [18, 19]. Similarly, FastSpeech 2.0 in TTS jointly learns prosodic parameters from text and speech [20]. Leveraging such rich representations can enhance prominence detection by integrating both acoustic and linguistic cues, addressing the limitations of handcrafted features.

Proposed work: We address two key challenges in word prominence detection: (1) modeling suprasegmental dependencies at the syllable level to capture fine-grained stress patterns instead of relying on word-level aggregates, and (2) exploring learned embeddings from FastSpeech 2.0's variance adaptor and wav2vec 2.0 to integrate both acoustic and linguistic information, moving beyond handcrafted prosodic features. For experiments, we consider non-native German (GER) and Italian (ITA) speakers of English from the ISLE corpus [21]. The results show that our syllable-dependency-based approach using wav2vec 2.0 features outperforms the unsupervised baseline (n-gram-based) by 19.83% & 24.01% and the supervised baseline (SVM) by 4.74% & 5.73% in GER & ITA, respectively.

2. Dataset

In this study, we use the ISLE corpus [21] for our experiments, which consists of 7,834 speech utterances from 46 non-native English learners (23 German (GER) and 23 Italian (ITA)). The dataset is manually annotated for word-level prominence markings for the first time. A team of five linguists provides syllable-level prominence labels, which are then manually verified. To derive syllable transcriptions, we use the P2TK [22] syllabification tool, while word boundaries are obtained from phone transcriptions through the forced alignment process using Kaldi ASR toolkit [23]. This ensures that each utterance contains precise word- and syllable-specific timestamps along with corresponding prominence labels (0 for non-prominent and 1 for prominent words). Adopting a word-centric unit of analysis, we transcribe prominence markers at the word level, ensuring accurate identification and consolidating them into a binary classification. The dataset was divided into training and test sets, maintaining a balanced distribution of speaker nativity, age, gender, and language proficiency, following the methodology in [21]

3. Methodology

Figure 1 presents the entire pipeline of the proposed approach. It consists of three main stages: Feature Extraction, Syllable-Level Feature Aggregation, and Modeling for Prominence Detection.

3.1. Feature Extraction

Given a speech utterance consisting of multiple words, we extract two different sets of features. Instead of relying on heuristic-based acoustic features, we aim to leverage representations from self-supervised learning (SSL) models, and models that learn prosody-aware embeddings, such as FastSpeech 2.0. These models offer a more structured representation that captures speech characteristics beyond heuristic-based features.

1. FastSpeech 2.0 (Phoneme-Level Features): In this work, we employ the FastSpeech 2.0 [20] variance adaptor for feature extraction, motivated by its effectiveness in[24], which has shown promising results. Given its ability to extract prosodic embeddings for duration, pitch, and energy, we use it to obtain a detailed representation of prosody variations. The variance adaptor learns the association between phonemes and prosodic characteristics. During training, it utilizes ground-truth duration, pitch, and energy values to train predictors for each component, ensuring that the extracted embeddings capture real-world prosody variations. We extract phoneme-level embeddings for pitch, duration, and energy, where each phoneme is represented by three separate 256-dimensional embeddings, one for each prosodic component. To obtain these embeddings, we fine-tune a pretrained FastSpeech 2.0 model, originally trained on the LJSpeech dataset, on the ISLE corpus for 12 epochs. This ensures that the variance adaptor learns prosody representations directly from speech, leading to more accurate embeddings. Since the pronunciation lexicon in FastSpeech 2.0 is based on native English speech, it may introduce errors when processing non-native speech. To address this, we dynamically modify the lexicon using manual phonetic transcriptions from the ISLE corpus, ensuring more accurate phoneme representations.

2. wav2vec 2.0 (Frame-Level Features) Self-supervised learning models like wav2vec 2.0 [18] and HuBERT [19] have shown strong performance in capturing speech representations without manual labeling. Among these, wav2vec 2.0 is widely used across speech tasks for its ability to encode spectral, phonetic, speaker, and semantic information directly from raw audio [25, 26]. These features have been effective in applications such as speech recognition and prosodic analysis. Since wav2vec 2.0 features have been successfully used for syllable stress detection in prior studies [27, 28], we incorporate them in this work to capture syllable dependencies, which are crucial for word-level prominence detection. By leveraging these representations, we aim to analyze how well they encode prominence patterns in speech. For our experiments, we extract 768-dimensional frame-level embeddings from wav2vec 2.0.

3.2. Syllable-Level Feature Averaging

The extracted features are averaged over each syllable segment to obtain syllable-level representations. Our approach is motivated by the need to incorporate syllable-level prominence dependencies for accurate word prominence detection. Since prominence manifests at the syllable level and influences word prominence [29], particularly in the non-native context, we align phoneme-level embeddings from FastSpeech 2.0 and frame-level features from wav2vec 2.0 to syllable boundaries using syllable time-stamp transcriptions. By averaging the extracted features within each syllable, we ensure consistency in representation, allowing the model to effectively capture prominence patterns at the syllable level before leveraging them for word-level.

3.3. Joint Syllable-Word Prominence Model (JSWPM)

The processed syllable-level features are passed through a Bidirectional LSTM (BLSTM) network to model sequential dependencies within a word. Our primary objective is word prominence detection, where we classify a word as prominent (1) or non-prominent (0). To enhance word prominence detection, we introduce an auxiliary task of syllable prominence detection, which helps the model learn prominence relationships between syllables within a word. Since word prominence emerges from the prominence of its syllables, this auxiliary task provides

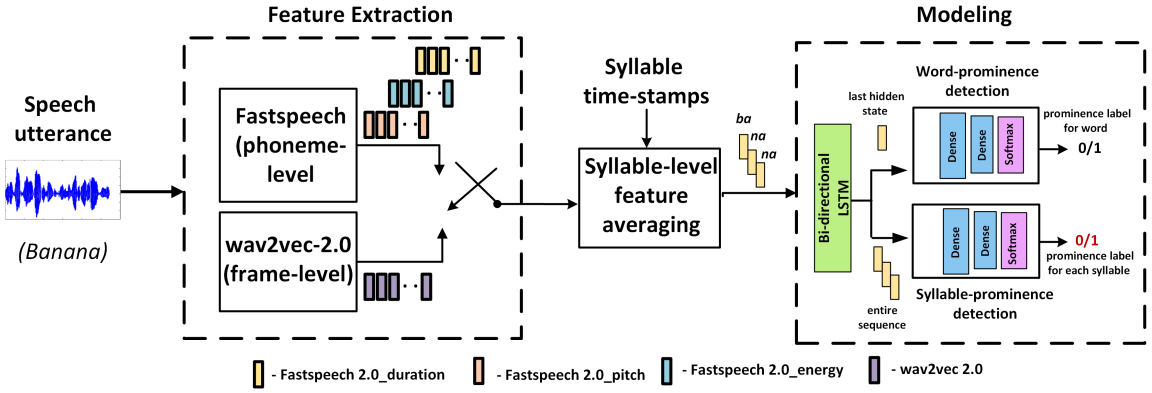


Figure 1: *Block diagram of proposed approach for word prominence detection*

finer-grained supervision, improving the model's ability to capture prominence structures. The model processes syllable sequences at the word level, ensuring all syllables of a given word are analyzed together. The BLSTM encodes the entire syllable sequence, and its outputs are used for two tasks:

1. Word-Prominence Detection (Primary Task): The model predicts whether the entire word is prominent (1) or not (0). For this, only the last hidden state from the final BLSTM layer is used, as it captures the accumulated prominence information across the syllable sequence. This representation is passed to a feedforward network with three dense layers (16, 4, and 1 neurons).

2. Syllable-Prominence Detection (Auxiliary Task): The goal is to predict prominence labels at the syllable level. Instead of taking only the last hidden state, the entire BLSTM output sequence is passed to a separate dense network with the same architecture (16, 4, 1 neurons). This ensures that prominence patterns at the syllable level are learned, which in turn supports the primary word prominence detection task.

Joint Training with Multi-Task Learning: Since our goal is to enhance word prominence detection by incorporating syllable-level prominence, we train the model using a multi-task learning framework. This involves computing separate losses for both tasks and jointly optimizing them. A categorical cross-entropy (CCE) loss is computed for both the word prominence and syllable prominence predictions. The total loss L is formulated as:

$$L = \lambda \sum_{i=1}^{N_w} \text{CCE}(\hat{y}_i^w, y_i^w) + (1-\lambda) \sum_{j=1}^{N_s} \text{CCE}(\hat{y}_j^s, y_j^s) \quad (1)$$

N_w represents the number of word-level samples, where $\hat{y}_i^w$ and y_i^w are the predicted and ground-truth word prominence labels, respectively. N_s represents the number of syllable-level samples, where $\hat{y}_j^s$ and y_j^s are the predicted and ground-truth syllable prominence labels, respectively. λ is a weighting factor that balances the contribution of the two tasks. Gradients are computed for both tasks and backpropagated through the shared BLSTM network, ensuring that the model learns from both objectives simultaneously. This allows the BLSTM to refine its representations by incorporating syllable-level dependencies while optimizing for word prominence detection.

The rationale behind this approach is that prominence is inherently structured at the syllable level, and learning syllable prominence helps in modeling word prominence more effectively. By sharing the BLSTM backbone and optimizing a joint loss, the model benefits from additional supervision, especially in cases where word prominence labels are scarce.

3.4. Word-Only Prominence Model (WOPM)

In scenarios where explicit syllable prominence labels are unavailable, we explore an alternative modeling approach. Instead of leveraging syllable-level prominence detection, we use a sequence of aggregated syllable-level features of a word as input to an BLSTM model. The classifier is trained solely for word-level prominence detection, using only the word-level loss. While this model does not explicitly incorporate syllable prominence, it still captures dependencies and variations across syllables within a word. By processing syllable-level sequences, the BLSTM learns patterns that contribute to better word prominence detection, even without explicit syllable prominence supervision. Figure 1 illustrates this alternative modeling approach, which follows the same word-prominence detection module as in the joint model but excludes the syllable-prominence detection module.

4. Experimental setup

We conduct experiments separately on German (GER) and Italian (ITA) datasets. Our study considers five feature sets: (1) Knowledge-based features (KBF), which include statistical measures of f0, energy, and duration, {2) pitch, 3) energy, and 4) duration embeddings extracted from FastSpeech 2.0}, and 5) embeddings from wav2vec 2.0. The FastSpeech 2.0 embeddings have a dimensionality of 512, while wav2vec 2.0 embeddings have a dimensionality of 768 at the phoneme and frame levels, respectively. We adopt the unsupervised n-gram model [4] as our baseline, as it captures sequential dependencies for word prominence detection, aligning with our sequence-dependent approach. Following prior work, we first extract acoustic features such as F0, duration, energy, and spectral tilt. Frame-level features are then quantized using k-means clustering, and the n-gram model is applied to the resulting cluster IDs to predict the next frame's probability, where lower scores indicate prominence. These scores are averaged across all frames within a word, and a threshold is applied to classify it as prominent or non-prominent.

In addition to the unsupervised approach, we consider and implement several supervised models for a fair comparison with our proposed method. These include Support Vector Machines (SVM) and Deep Neural Networks (DNN), where the input features are aggregated at the word level by averaging across frames. SVM used with kernel as radial basis function with C = 100. DNN architecture details: The DNN model comprises five dense layers with hidden units configured as follows: 64, 32, 16, 4, and 1. It employs the Rectified Linear Unit (ReLU) activation function [30].

Table 1: *F1-scores of n-gram model for word prominence detection*

	GER	ITA
F0	55.39	72.25
Energy	58.34	72.65
Spectral Tilt	55.33	71.12
F0+Energy+SpectralTilt	56.8	72.55

Proposed model architecture: The model consists of a single-layer bi-directional BLSTM with 512 hidden units. Two fully connected layers are utilized for word-level and syllable-level classification, each producing two logits. A softmax activation function is applied to both outputs, with a loss balancing factor of $\lambda = 0.5$. The small architecture prevents overfitting given the limited training data while efficiently capturing sequential dependencies among syllables within a word. We use the F1-score as the evaluation metric. All models are optimized using the Adam optimizer [31] with categorical cross-entropy as the loss function. The number of layers and hidden units are selected based on optimal performance on the validation dataset.

5. Results and Discussion

In this section, we first present the results of our baseline (n-gram). Next, we report the results of all supervised approaches from the literature, which are sequence-independent, and compare them with our proposed sequence-dependent approach. In the proposed approach, we conduct experiments both without explicitly using syllable prominence markings, which we refer to as WOPM, and with syllable prominence markings, which we refer to as JSWPM, as described in the methodology. Table 1 presents the results of the baseline, while Table 2 compares the results of all supervised approaches with the proposed approach. We evaluate performance separately on the GER and ITA datasets using F1-scores. We consider five different feature sets in our experiments: Knowledge-based features (KBF), three different embeddings (pitch, energy, and duration) extracted from the variance adaptor of FastSpeech, and wav2vec 2.0 embeddings.

The results indicate that supervised models significantly outperform the unsupervised n-gram baseline. The highest performance achieved by the n-gram model (58.34 F1-score for GER and 72.65 for ITA) is substantially lower than any of the supervised approaches, reinforcing the limitations of handcrafted prosodic features and statistical modeling in capturing prominence patterns. Among the supervised methods, the proposed JSWPM model consistently achieves the highest F1-scores across all feature sets, demonstrating the benefits of incorporating sequential dependencies at the syllable level. Comparing knowledge-based heuristic features (KBF) with features extracted from FastSpeech 2.0 and wav2vec 2.0, we observe that the latter consistently achieve better performance with the proposed approach, highlighting the advantage of learned representations over heuristics-based features. Among the FastSpeech 2.0-derived embeddings (F_pitch, F_energy, and F_duration), F_pitch and F_energy consistently outperform F_duration in both GER and ITA. In GER, JSWPM achieves the highest F1-scores with F_pitch (75.03) and F_energy (74.94), outperforming SVM by 6.84% and 4.76% and DNN by 3.82% and 7.98%, respectively. A similar trend is observed for ITA, where JSWPM again performs best with F_pitch (83.29), surpassing F_duration (82.57). These results highlight that pitch and energy variations are more effective for prominence detection, and JSWPM consistently benefits the most from these features compared to the supervised baseline models (SVM and DNN).

Table 2: *Comparison of average F1-scores for word prominence detection on GER and ITA datasets using different ML classifiers (SVM and DNN) and the proposed (WOPM and JSWPM) models. 'F' refers to embeddings extracted from FastSpeech 2.0.*

		SVM	DNN	WOPM	JSWPM
GER	**KBF**	73.43	73.24	-	**-**
	F_Pitch	68.19	72.27	74.69	**75.03**
	F_Duration	66.91	66.64	74.12	**74.61**
	F_Energy	70.18	69.42	73.45	**74.94**
	wav2vec-2.0	74.77	76.47	**77.38**	**78.17**
ITA	**KBF**	80.16	80.83	-	**-**
	F_Pitch	81.37	80.39	82.47	**83.29**
	F_Duration	78.88	77.24	82.1	**82.57**
	F_Energy	81.44	81.66	82.62	82.85
	wav2vec-2.0	85.08	85.43	86.06	86.56

Considering wav2vec 2.0 embeddings, for GER, JSWPM achieves 78.17 with wav2vec 2.0, marking a 4.56% improvement over SVM (74.77) and 2.22% over DNN (76.47). For ITA, JSWPM achieves 86.56, showing a 1.74% gain over SVM (85.08) and 1.31% over DNN (85.43). These improvements again highlight the effectiveness of the proposed approach. Further, across all architectures, models trained on wav2vec 2.0 embeddings outperform those using FastSpeech 2.0-derived features, demonstrating the effectiveness of self-supervised representations in capturing prominence cues. In GER, wav2vec 2.0 improves performance by approximately 3-5% over FastSpeech 2.0 features, while in ITA, the gains are around 3%. These results suggest that wav2vec 2.0 provides a more comprehensive representation of prominence by capturing both acoustic and linguistic cues more effectively than FastSpeech 2.0, which primarily encodes prosody-driven speech synthesis features.

Across different classifiers, the proposed WOPM and JSWPM approaches outperform SVM and DNN in most cases, with JSWPM consistently achieving the highest accuracy. While WOPM leverages word-level representations, JSWPM further incorporates syllable-level prominence modeling, leading to improved performance. The results indicate that explicitly capturing suprasegmental dependencies enhances prominence detection. Although the performance gains of JSWPM over WOPM are modest, its ability to model syllable-level variations suggests potential for further improvements with more data, reinforcing the importance of syllable-level prominence information in word prominence detection.

6. Conclusion

We demonstrate the effectiveness of syllable-level modeling for word prominence detection, showing that capturing suprasegmental dependencies enhances performance. Our experiments on English from non-native speakers (German and Italian) speech from the ISLE corpus show that the proposed JSWPM approach consistently outperforms WOPM and traditional classifiers such as SVM and DNN. Specifically, JSWPM with wav2vec 2.0 improves over SVM by 4.74% for GER and by 5.73% for ITA. Additionally, self-supervised embeddings from wav2vec 2.0 outperform FastSpeech 2.0-derived features, capturing both acoustic and linguistic prominence cues. Since WOPM and JSWPM perform on par, even without explicit syllable prominence labels, our findings suggest that prominence detection can be effectively modeled without direct supervision at the syllable level. This is particularly useful in low-resource scenarios. Our study highlights the significance of syllable-level representations in automatic prominence detection and offers insights for spoken language understanding.

7. References

[1] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment: Studies presented to Gösta Bruce*. Springer, 2000, pp. 89–127.

[2] A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[3] S. Kakouros, O. Räsänen, and P. Alku, "Comparison of spectral tilt measures for sentence prominence in speech—effects of dimensionality and adverse noise conditions," *Speech Communication*, vol. 103, pp. 11–26, 2018.

[4] S. Kakouros and O. Räsänen, "Automatic detection of sentence prominence in speech using predictability of word-level acoustic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] T. Mishra, V. K. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features." in *Interspeech*, 2012, pp. 1864–1867.

[6] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 2, pp. 690–701, 2007.

[7] J. Chi, M. de Seyssel, and N. Schluter, "The role of prosody in spoken question answering," *arXiv preprint arXiv:2502.05389*, 2025.

[8] M. Levy, *Computer-assisted language learning: Context and conceptualization*. Oxford University Press, 1997.

[9] K. Imoto, M. Dantsuji, and T. Kawahara, "Modelling of the perception of english sentence stress for computer-assisted language learning," in *Sixth International Conference on Spoken Language Processing*, 2000.

[10] S. Kakouros, O. Räsänen, and P. Alku, "Evaluation of spectral tilt measures for sentence prominence under different noise conditions," in *Interspeech*. International Speech Communication Association (ISCA), 2017, pp. 3211–3215.

[11] K. Sabu and P. Rao, "Detection of prominent words in oral reading by children," *Proceedings of speech prosody, Poznan, Poland*, pp. 314–318, 2018.

[12] J. M. Brenier, D. M. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features." in *Interspeech*, 2005, pp. 3297–3300.

[13] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive science*, vol. 40, no. 7, pp. 1739–1774, 2016.

[14] S. Kakouros, J. Pelemans, L. Verwimp, P. Wambacq, and O. Räsänen, "Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence." in *Interspeech*, 2016, pp. 1074–1078.

[15] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning." in *Interspeech*, 2015, pp. 3066–3070.

[16] A. Cutler and D. J. Foss, "On the role of sentence stress in sentence processing," *Language and speech*, vol. 20, no. 1, pp. 1–10, 1977.

[17] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–937.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[19] 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.

[20] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[21] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in Proceedings of LREC: Language Resources and Evaluation Conference, vol. 2. European Language Resources Association, 2000, pp. 957–964.

[22] J. Tauberer, "P2tk automated syllabifier," 2008.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[24] A. Mondal, R. S. Bharadwaj, J. Mallela, A. K. Vuppala, and C. Yarra, "A preliminary analysis of automatic word and syllable prominence detection in non-native speech with text-to-speech prosody embeddings," *arXiv preprint arXiv:2412.08283*, 2024.

[25] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.

[26] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[27] A. de la Fuente and D. Jurafsky, "A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models," in *Interspeech 2024*, 2024, pp. 1290–1294.

[28] S. H. Aluru, J. Mallela, and C. Yarra, "Post-net: A linguistically inspired sequence-dependent transformed neural architecture for automatic syllable stress detection," in *Proc. Interspeech 2024*, 2024, pp. 3340–3344.

[29] D. Arnold, B. Möbius, and P. Wagner, "Comparing word and syllable prominence rated by naıve listeners," in *Proceedings of INTERSPEECH*, 2011.

[30] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.