Interspeech 2024
1-5 September 2024, Kos, Greece

Post-Net: A linguistically inspired sequence-dependent transformed neural architecture for automatic syllable stress detection

Sai Harshitha Aluru[1], *Jhansi Mallela*[2], *Chiranjeevi Yarra*[2]

[2]Speech Processing Lab, LTRC, IIIT Hyderabad, India
[1]Department of Artificial Intelligence, Vidya Jyothi Institute of Technology, India
saiharshitha.192@gmail.com, jhansi.mallela@research.iiit.ac.in,
chiranjeevi.yarra@iiit.ac.in

Abstract

Automatic syllable stress detection methods typically consider syllable-level features as independent. However, as per linguistic studies, there is a dependency among the syllables within a word. In this work, we address this issue by proposing a Post-Net approach using Time-Delay Neural Networks to exploit the syllable dependency in a word for stress detection task. For this, we propose a loss function to incorporate the dependency by ensuring only one stressed syllable in a word. The proposed Post-Net leverages the existing SOTA sequence-independent stress detection models and learns in both supervised and unsupervised settings. We compare the Post-Net with three existing SOTA sequence-independent models and also with sequential model (LSTMs). Experiments conducted on ISLE corpus show the highest relative accuracy improvement of 2.1% and 20.28% with the proposed Post-Net compared to the best sequence-independent SOTA model in supervised and unsupervised manners, respectively.

Index Terms: Syllable stress detection, Time-Delay Neural networks, Post-Net

1. Introduction

Syllable stress refers to the emphasis or prominence placed on a syllable within a word. A word is usually a sequence consisting of stressed and unstressed syllables. Stressed syllables[1] often receive greater articulatory effort and exhibit differences in characteristics such as loudness, pitch, and duration compared to unstressed syllables [2]. Incorrectly placing stress on syllables can alter the meaning of a word and lead to confusion in both human-human and human-machine interactions. This issue is particularly prevalent among second language (L2) learners, who often carry over stress patterns from their native language. To aid L2 learners, computer-assisted language learning (CALL) systems play a crucial role in detecting syllable stress errors [3]. Automatic syllable stress detection forms a key component of CALL systems.

Typically, syllable stress detection is a two-class classification problem. For this, in the literature, various acoustic features reflecting energy, pitch, and duration were computed [4, 5, 6, 7]. Despite advancements in heuristics-based features, stress detection remains challenging due to the varying characteristics of syllables across different words. To address this, deep learning models have been proposed to learn the complex and non-linear relationships across the features. Shahin et al. [8] demonstrated the effectiveness of convolutional neural networks (CNN) in stress detection using prosodic and spectral features. Song et al. [9] employed deep belief networks for lexical stress classification, outperforming traditional Gaussian Mixture Models. Tian et al. [10] utilized attention-based neural networks and bidirectional LSTMs with MFCCs, energy, and pitch features. Ruan et al. [11] applied a transformer network for stress detection. In [12], a novel approach was introduced, emphasizing the necessity of representation learning for stress detection. This approach involves jointly optimizing variational autoencoders and DNNs to effectively model the intricate details and interdependencies among the stress features within a syllable.

In all the works, the stress detection is modeled considering each syllable-level feature independently. However, according to the phonology of English, each word typically carries only one primary stressed syllable. Often, in the literature, this constraint was incorporated as a separate post-processing step after the classification considering the probabilities from the classifier to achieve a higher performance. This is because the classifier could not learn the sequence dependency posed by the constraint. Thus, we believe that the classifier learned with the sequence dependency constraint could result in better performance than that without the sequence dependency.

To incorporate sequence dependency constraint, we propose a Post-Net approach with a custom loss function inspired by the constraint. The Post-Net considers a time-delay neural network to incorporate the constraint and extends the deep learning (DL) based supervised syllable sequence independent models for capturing sequence dependency. In the Post-net, a masking strategy is used to incorporate the constraint for all varying-length sequences. We also show that the proposed Post-net can be extended to unsupervised settings.

In this work, the proposed Post-net is applied on three existing state-of-the-art supervised deep-learning based stress detection models and in an unsupervised setting using a deep neural network. Experiments are conducted on non-native English collected from German and Italian speakers. We explore two different sets of features for stress detection: 1) state-of-the-art heuristics-based acoustic and context features, and 2) self-supervised feature representations (Wav2Vec-2.0). We consider the three different state-of-the-art non-sequential models as baselines [12]. Also, we compare the performance of proposed Post-Net model with a sequential model (LSTMs). From the experimental results, we observe the highest relative improvement of 3% & 2.13% and 6.4% & 4% in terms of classification accuracies under supervised and unsupervised settings in both German & Italian, respectively over baseline.

2. Dataset

For experiments, we consider the ISLE corpus [13]. It consists of 7834 spoken utterances from 46 non-native English speakers out of which 23 are German (GER), and 23 are Italian (ITA). Each speaker contributed 160 utterances. We obtained phoneme transcripts for all utterances using forced-alignment process. A group of five linguists manually verified these alignments to ensure they accurately represented the speakers' pro-

10.21437/Interspeech.2024-2400

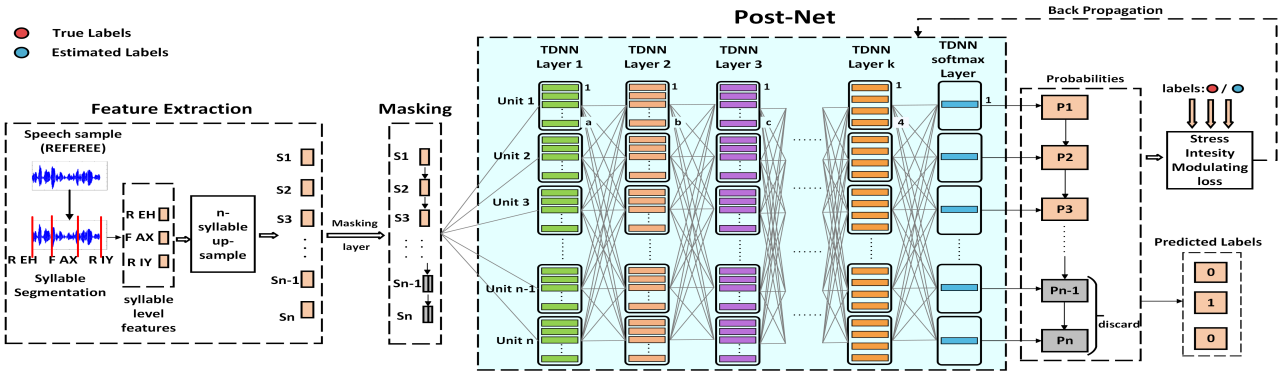

Figure 1: *Block Diagram of the proposed Post-Net approach*

nunciation. Using P2TK syllabification software [14], we generated syllable transcripts from phoneme transcripts. Stress labels are annotated for each syllable by the same team of linguists. It is ensured that each word contains only one stressed syllable. We consider only polysyllabic words which resulted in 12,388 stressed and 16,005 unstressed syllables. Train and test splits for both the German and Italian datasets are balanced in terms of factors such as nativity, age, sex, and proficiency of the speakers, as detailed in literature [13].

3. Proposed Approach

The proposed Post-Net [1] approach considers the temporal sequence of syllable-level features to learn their dependencies at the word level. The Post-Net is designed such that it can be integrated with any deep learning based sequence-independent models for leveraging the dependency. We discuss the Post-Net integration process considering three sequence-independent models with the help of block diagrams in Figures 1 & 2 and the following subsections: 1) Data processing, 2) Masking, 3) Post-Net modeling, and 4) Stress Intensity Modulating Loss.

3.1. Data processing:

Given a speech utterance of a word with r syllables, first, syllable segments are obtained using their time-aligned boundaries. Then, features of d-dimension are extracted for each syllable. After feature extraction, a set of $n-r$ d-dimension features are concatenated to the extracted features to make a total of n d-dimension features. This is to make all words have an equal number of syllables for satisfying the constraint posed by the generic sequence-dependent models' training. In this work, we consider n as the highest number of syllables among all the words in the dataset. In this corpora, the value of n is 5. The d elements in the $n-r$ concatenated features are assigned to -1. Though we assigned -1, any real value can be considered. This is because these assigned values will be neglected in the further steps with the masking strategy.

3.2. Masking:

Masking layers are typically used in DL-based modelling for efficiently handling variable-length sequences by ignoring irrelevant or padded elements during training. Thus, in this work, we use the masking layer of n elements to conceal the concatenated $n-r$ d-dimensional syllable features to avoid their contribution in the training process. The resulting sequence of syllable features after masking is fed to the Post-Net model.

3.3. Post-Net modelling:

In our approach, we utilize Time-Delay Neural Networks (TDNNs) to construct the Post-Net model. TDNN is a neural network architecture with multiple hidden layers specifically designed for handling sequential data and popularly used in handling speech-based applications including speech recognition [15]. Each layer in TDNN processes input features across different time steps or delays, enabling the network to capture temporal dependencies present in the input data sequence [16]. We believe that the time-delay concept inherent in TDNNs aids in learning the sequential dependency within the r d-dimension syllable level features of a word, as it allows for sequential feature grouping.

As depicted in the block diagram (see Figure 1), the Post-Net model consists of k TDNN layers, each comprising n time-delay units and $k+1$th layer with softmax as an activation function. For a given r syllabic word, first, we extract the stress probabilities for n concatenated features from the softmax layer. Then we exclude the probabilities belonging to $n-r$ masked features, and finally, we obtain probabilities for r syllables.

Integration process: This process is achieved by initializing the TDNN units in each layer of the Post-Net with already trained sequence-independent stress detection models, referred to as Pre-Net models. We utilize three different state-of-the-art [12] sequence-independent syllable-level stress detection models as Pre-Net models. The choice of these models is based on the following type of representation learning mechanisms; i) Task specific representations as observed in simple DNN, ii) Distribution specific representations as observed in Variation autoencoder (VAE) followed by task-specific representations with DNN; referred to as VAE_DNN, and iii) Jointly learnt distribution, and task-specific representations which can be obtained by jointly trained VAE and DNN; referred as J_VAE_DNN.

After the integration process, the Post-Net models obtained from the three different sequence independent models, DNN, VAE_DNN and J_VAE_DNN, referred to as PN_DNN, PN_VAE_DNN and PN_J_VAE_DNN. The detailed descriptions of these three Post-Net are given below.

i) PN_DNN: Figure 1 illustrates the architecture of the Post-Net using DNN with k layers as the Pre-Net. In every layer of PN_DNN, each TDNN unit is initialized with the parameters (hidden units and weights) of the respective layer in DNN.

ii) PN_VAE_DNN: Figure 2.a illustrates the working flow of VAE_DNN without the dotted line and Figure 2.b illustrates the corresponding Post-Net model, PN_VAE_DNN with n sequence. As shown in the figure, the input to VAE_DNN is at syllable-level (X) while it is sequence (X_n) in PN_VAE_DNN. Unlike DNN, VAE_DNN comprises encoder and decoder networks. Hence, in PN_VAE_DNN, two Post-Net models: Post-Net1 and Post-Net2, are added mirroring the encoder and decoder architectures, respectively. The weights of each unit in layers of Post-Net1 and Post-Net2 are initialized with respective layers in encoder and decoder models of VAE.

iii) PN_J_VAE_DNN: Figure 2.a illustrates the working

[1] https://github.com/jhansimallela/Interspeech2024_PostNet

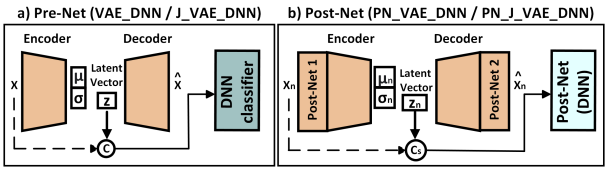

Figure 2: *Block Diagram of Pre-Net(a) and Post-Net(b) models based on VAE_DNN & J_VAE_DNN*

flow of J_VAE_DNN with the dotted line and Figure 2.b illustrates the corresponding Post-Net model, PN_J_VAE_DNN with n length sequence. Similar to PN_VAE_DNN, we build Post-Net models for PN_J_VAE_DNN and are initialized with the weights of the encoder and decoder of J_VAE_DNN.

Therefore, the proposed approach involves two main steps: i) Training the Pre-Net models for syllable stress detection at the syllable level, ii) Initializing the weights of each TDNN unit in the Post-Net model with the corresponding layers of the Pre-Net models and fine-tuning the Post-Net model with the sequence of syllable-level features in a sequence-dependent manner. We fine-tune the Post-Net model using a custom loss function called the stress intensity modulating loss. This loss function utilizes the predicted probabilities from the Post-Net model and custom labels, which vary for supervised and unsupervised cases.

3.4. Stress intensity modulating loss (SIML):

The objective of the proposed custom loss function, SIML is to maximize the likelihood of one syllable as stressed and all remaining syllables as unstressed in a word. SIML is computed between the sequence of probabilities given by the last layer of the Post-Net model and pseudo labels. Pseudo labels differ in in supervised and unsupervised settings:

Unsupervised setting: In this setting, we assign pseudo labels based on the probabilities predicted by the Post-Net for the sequence of syllables in each word. The syllable with the highest probability of being stressed is given a pseudo label as '1' and others as '0'. This pseudo-labeling strategy allows us to establish a form of supervision without explicit labels.

Supervised setting: In this setting, groundtruth stress labels are considered as the pseudo labels.

In both supervised and unsupervised settings, aligning with the objective of SIML, we have observed that binary cross-entropy loss calculated across the sequence of predicted probabilities and pseudo stress labels effectively fulfills this objective.

4. Experimental Setup

In this study, we experiment with the following two different sets of features. 1) Knowledge-based features - Following the work in [17], we consider 19-dimensional acoustic features along with 19-dimensional context features. 2) Self-supervised representations (Wav2Vec-2.0) - These features have been outperforming several heuristics-based features in many speech applications [18, 19]. We utilize these features (768-dimensional), which encompass phonetic, spectral, temporal, and contextual information. We hypothesize that these features capture stress-related information, thereby enhancing stress detection performance. Typically these are extracted at the frame level, thus to obtain syllable-level features, we average the frame-level features across all the frames within the syllable segment. We first train the Pre-Net models (DNN, VAE_DNN, and J_VAE_DNN) at the syllable level. Further, the weights of the Pre-Net model layers are used for the initialization of Post-Net architecture for fine-tuning the sequence model at word-level. We consider all three Pre-Net models as baselines in this study.

Baseline with a generic sequential network (LSTMs): In addition to comparing the Post-Net with Pre-Net models, which are independent of words, we also conduct experiments using a sequential model built based on LSTMs and compare its performance with our proposed Post-net models. Unlike our proposed approach, the LSTM-based approach could not integrate the Pre-Net model. However, both the Post-Net approach and the LSTM-based approach utilize the same input: a sequence of syllable-level features.

Architecture details: The DNN model consists of five dense layers [Hidden units: 64,32,16,4,1] with a Rectified Linear Unit (ReLU) [20] activation function and Adam [21] as the optimizer. The VAE model has one hidden layer in the encoder and decoder, using the ReLU activation function. The optimal parameters for both models are selected based on maximizing performance on the validation set. The three Post-Net (PN_DNN, PN_VAE_DNN, and PN_J_VAE_DNN) mirrors the respective Pre-Net model architecture, DNN, VAE_DNN, or J_VAE_DNN. In the LSTM-based model, we consider three LSTM layers [number of cells in each layer: 64, 32, 16] followed by two dense layers [hidden units: 4,1] with a softmax activation function in the last layer. We conduct all the experiments in a five-fold cross-validation setup under the following three scenarios: **Matched:** Individual models are trained with GER and ITA train sets, and then tested separately on the GER and ITA test sets, respectively. **Combined:** A single model is trained using combined data from both GER and ITA, and tested separately on the GER and ITA test sets. **Cross:** Models are trained with GER (ITA) train set and then tested on the ITA (GER) test set.

5. Results and Discussion

We analyse the performance of the proposed Post-Net with the respective Pre-Net using knowledge-based (A+C) and Wav2Vec 2.0 features in both supervised and unsupervised settings under the three scenarios (matched, combined, and cross) for GER and ITA data. Also, we analyse visualization of the learned representations from Post-Net and Pre-Net using t-SNE [22]. At the end, we compare the Post-Net performance with the LSTM.

5.1. Comparison with Pre-Net using knowledge-based features

Table 1 presents the classification accuracies (F1-scores in brackets) under three different scenarios for both GER and ITA. From the results, it is observed that in all the scenarios of both GER and ITA, the proposed Post-Net using TDNN outperforms the baselines (Pre-Net models). This suggests that the Post-Net approach captures the sequential information and dependencies among the syllable sequence in a word. Among the matched, combined, and cross scenarios, The highest accuracies are resulted in the combined scenario, which are found to be 94.79% and 94.9% for GER and ITA with PN_J_VAE_DNN.

5.2. Comparison with Pre-Net using Wav2Vec-2.0 features

Table 2 presents the results of DNN vs PN_DNN, VAE_DNN vs PN_VAE_DNN and J_VAE_DNN vs PN_J_VAE_DNN on Wav2Vec-2.0 features in matched, combined, and cross scenarios. Similar to Table 1, it is observed that the proposed approach is outperforming the baseline models in every case with significant improvements. For PN_J_VAE_DNN, the highest accuracy of 95.36% and 95.5% is obtained on GER and ITA, respectively. Compared to the performance with knowledge-based

Table 1: *Accuracies and F1-scores (in brackets) of Pre-Net (DNN, VAE_DNN, J_VAE_DNN) and Post-Net (PN_DNN, PN_VAE_DNN, PN_J_VAE_DNN) considering knowledge-based features under three different scenarios.*

knowledge-based		DNN	PN_DNN	VAE_DNN	PN_VAE_DNN	J_VAE_DNN	PN_J_VAE_DNN
GER	Matched	92.81 (91.81)	**92.97 (91.23)**	90.98 (89.94)	**92.15 (91.59)**	93.36 (92.42)	**93.5 (92.38)**
	Combined	92.86 (91.76)	**93.4 (92.63)**	93.39 (92.50)	**94.56 (93.22)**	93.89 (93.06)	**94.79 (93.58)**
	Cross	88.44 (86.85)	**89.19 (88.71)**	86.70 (85.02)	**87.68 (86.47)**	88.60 (86.73)	**90.20 (89.06)**
ITA	Matched	91.33 (90.07)	**92.37 (90.95)**	90.20 (88.98)	**91.09 (90.47)**	92.54 (91.50)	**93.33 (92.25)**
	Combined	92.71 (92.23)	**93.70 (92.89)**	92.84 (91.80)	**94.53 (93.17)**	94.68 (93.52)	**94.9 (93.74)**
	Cross	91.02 (89.72)	**91.57 (90.44)**	87.28 (86.88)	**88 (87.12)**	90.80 (90.50)	**91.63 (90.37)**

features (Table 1), the accuracy obtained using Wav2Vec 2.0 is more with a relative improvement of 1.34%, 0.61%, 3.05% and 1.59%, 0.63%, 2.03% under matched, combined, cross of GER and ITA, respectively. This indicates the effectiveness of the Wav2Vec-2.0 features compared to the knowledge-based features.

Table 2: *Accuracies and F1-scores (in brackets) of Pre-Net (DNN, VAE_DNN, J_VAE_DNN) and Post-Net (PN_DNN, PN_VAE_DNN, PN_J_VAE_DNN) considering Wav2Vec-2.0 features under three different scenarios.*

Wav2Vec-2.0		DNN	PN_DNN	VAE_DNN	PN_VAE_DNN	J_VAE_DNN	PN_J_VAE_DNN
GER	Matched	94.05 (93.23)	**94.66 (93.60)**	92.25 (91.75)	**93.5 (92.34)**	94.38 (93.37)	**94.75 (93.86)**
	Combined	94.10 (94.43)	**95.21 (94.93)**	93.50 (92.60)	**94.62 (93.3)**	94.73 (93.81)	**95.36 (94.74)**
	Cross	90.75 (89.48)	**92.65 (91.41)**	88.91 (87.42)	**90.27 (89.05)**	92.3 (91.23)	**92.90 (91.85)**
ITA	Matched	93.61 (92.70)	**94.31 (93.60)**	92.12 (91.65)	**93.32 (92.3)**	93.63 (92.71)	**94.8 (93.49)**
	Combined	94.34 (93.98)	**95.35 (94.86)**	94.10 (93.20)	**95.18 (94.37)**	94.90 (94.28)	**95.50 (94.72)**
	Cross	92.10 (91.21)	**93.32 (92.04)**	90.30 (89.58)	**92.01 (89.65)**	92.15 (91.64)	**93.47 (92.65)**

5.3. Comparison under unsupervised setting

Table 3 presents the results of K-Means clustering and Post-Net approach in unsupervised setting with DNN architecture referred to as U_TDNN. In this case, the Post-Net approach does not use the Pre-Net model weights for initialization. From the results, it is clearly evident that the Post-Net approach is outperforming in all cases compared to K-Means with significant improvement. Similar to supervised setting, Wav2Vec-2.0 is performing better compared to knowledge-based. The highest accuracies obtained are 72.35% & 72.61% and 75.96% & 76.75% with a relative improvement of 10.69% & 12.7% and 10.42% & 13.05% compared to K-Means under the combined scenario with knowledge-based and Wav2Vec-2.0 in GER & ITA, respectively. These improvements prove that the Post-Net approach can capture the sequence dependencies well even in an unsupervised setting.

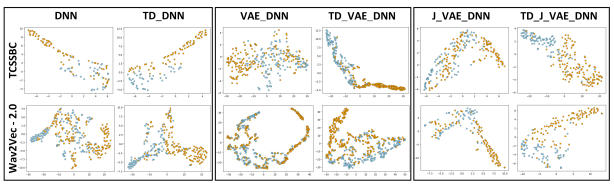

Figure 3: *t-SNE visualizations of learned representations from Pre-Net and Post-Net models (blue-stressed, orange-unstressed).*

5.4. t-SNE visualizations of Pre-Net and Post-Net models

Figure 3 showcases t-SNE visualizations of representations obtained from Pre-Net and Post-Net models using knowledge-based and Wav2Vec-2.0 features. From the plots, the key observations are as follows: i) Clearer distinction between stressed and unstressed classes is evident in Post-Net models compared to the Pre-Net models, supporting the hypothesis regarding the importance of learning sequential information, and ii) Wav2Vec-2.0 features demonstrate better separation in the learned representations, aligning with the trend observed in classification accuracies.

Table 3: *Accuracies and F1-scores (in brackets) of Kmeans vs Unsupervised Post-Net approach (U_TDNN) considering knowledge-based and Wav2Vec-2.0 features under three different scenarios.*

		knowledge-based		Wav2Vec-2.0	
		Kmeans	U_TDNN	Kmeans	U_TDNN
GER	Matched	64.66 (63.24)	**71.61 (70.05)**	65.28 (64.17)	**75.24 (74.39)**
	Combined	65.36 (65.84)	**72.35 (71.59)**	68.74 (67.62)	**75.96 (74.89)**
	Cross	63.99 (62.17)	**71.31 (69.98)**	64.16 (63.09)	**73.43 (72.17)**
ITA	Matched	59.76 (58.83)	**71.88 (70.23)**	64.05 (63.24)	**75.63 (74.51)**
	Combined	64.43 (63.09	**72.61 (71.44)**	67.89 (66.65)	**76.75 (75.03)**
	Cross	59.26 (58.03)	**71.05 (69.64)**	63.41 (62.7)	**74.3 (72.94)**

5.5. Comparison of Post-Net with LSTMs

Table 4 presents the performance results of Wav2Vec-2.0 features using LSTM and Post-Net models for both supervised and unsupervised cases. We present results only on Wav2Vec-2.0 features due to its consistent superiority over knowledge-based features as observed in previous evaluations. In the unsupervised setting, the proposed SIML loss function is used to train LSTM as proposed in Post-Net. Across all cases, it is observed that the Post-Net outperforms LSTMs, highlighting that the proposed Post-Net excels not only compared to Pre-Net models (non-sequential) but also in comparison with sequential models.

Table 4: *Accuracies and F1-scores (in brackets) of LSTM and Post-Net considering Wav2Vec-2.0 features under three different scenarios.*

		Supervised		Unsupervised	
		LSTM	PN_DNN	LSTM	U_TDNN
GER	Matched	93.12 (92.21)	**94.66 (93.6)**	70.71 (69.5)	**75.24 (74.39)**
	Combined	93.77 (92.53)	**95.21 (94.93)**	74.63 (73.17)	**75.96 (74.89)**
	Cross	89.98 (86.95)	**92.65 (91.41)**	70.24 (68.23)	**73.43 (72.17)**
ITA	Matched	93.143 (92.26)	**94.31 (93.6)**	72.73 (71.52)	**75.63 (74.51)**
	Combined	93.79 (92.7)	**95.35 (94.86)**	75.11 (74.16)	**76.75 (75.03)**
	Cross	91.369 (90.48)	**93.32 (92.04)**	73.64 (71.62)	**74.3 (72.94)**

6. Conclusion

In this study, we proposed a linguistically motivated syllable-sequence modeling approach, Post-Net (sequence-dependent) using TDNNs for automatic syllable stress detection. The proposed approach leverages the existing three different state-of-the-art syllable-level stress detection models. The proposed custom loss function in Post-Net facilitates modeling stress detection task in both supervised and unsupervised settings. We evaluated our proposed Post-Net models and compared their performance against the sequence-independent models and the sequence-dependent models with LSTMs. Also, we compared the performance of Post-Net model with K-Means in an unsupervised setting. Our experimental results on the ISLE corpus showed that the Post-Net model is able to learn the sequence dependency among the syllables and enhance the performance of stress detection task in both supervised and unsupervised settings.

7. References

[1] R. Goedemans, J. Heinz, and H. van der Hulst, *The study of word stress and accent: Theories, methods and data*. Cambridge University Press, 2018.

[2] A. Cutler and S. D. Isard, "The production of prosody," 1980.

[3] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[4] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in International Conference on Acoustics, Speech, and Signal Processing, vol. 1. IEEE, 2005, pp. 937–940.

[5] O. D. Deshmukh and A. Verma, "Nucleus-level clustering for word-independent syllable stress classification," *Speech Communication*, vol. 51, no. 12, pp. 1224–1233, 2009.

[6] J. Zhao, H. Yuan, J. Liu, and S. Xia, "Automatic lexical stress detection using acoustic features for computer assisted language learning," *Proc. APSIPA ASC*, pp. 247–251, 2011.

[7] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5845–5849.

[8] M. A. Shahin, J. Epps, and B. Ahmed, "Automatic classification of lexical stress in english and arabic languages using deep learning." in *Interspeech*, 2016, pp. 175–179.

[9] S.-H. Song and D. K. Kim, "Development of a stress classification model using deep belief networks for stress monitoring," *Healthcare informatics research*, vol. 23, no. 4, pp. 285–292, 2017.

[10] T. Xia, X. Rui, C. L. Huang, I. H. Chu, S. Wang, and M. Han, "An Attention Based Deep Neural Network for Automatic Lexical Stress Detection," in Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2019, pp. 1–5.

[11] Y. Ruan, X. Wang, H. Liu, Z. Ou, Y. Gao, J. Cheng, and Y. Qian, "An end-to-end approach for lexical stress detection based on transformer," *arXiv preprint arXiv:1911.04862*, 2019.

[12] J. Mallela, P. S. Boyina, and C. Yarra, "A comparison of learned representations with jointly optimized vae and dnn for syllable stress detection," in *International Conference on Speech and Computer*. Springer, 2023, pp. 322–334.

[13] 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 2000: Language Resources and Evaluation Conference, vol. 2*. European Language Resources Association, 2000, pp. 957–964.

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

[15] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Backpropagation*. Psychology Press, 2013, pp. 35–61.

[16] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *Interspeech*, 2015, pp. 3214–3218.

[17] C. Yarra, M. K. Ramanathi, and P. K. Ghosh, "Comparison of automatic syllable stress detection quality with time-aligned boundaries and context dependencies." in *SLaTE*, 2019, pp. 79–83.

[18] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for mispronunciation detection." in *Interspeech*, 2021, pp. 4428–4432.

[19] M. Kunešová and M. Řezáčková, "Detection of prosodic boundaries in speech using wav2vec 2.0," in *International Conference on Text, Speech, and Dialogue*. Springer, 2022, pp. 377–388.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

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

[22] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.