A Comparison of Learned Representations with Jointly Optimized VAE and DNN for Syllable Stress Detection

Jhansi Mallela(✉), Prasanth Sai Boyina, and Chiranjeevi Yarra

Speech Processing Lab, Language Technologies Research Center (LTRC),
International Institute of Information and Technology, Hyderabad 500032, India
jhansi.mallela@research.iiit.ac.in, prasanth.sai@students.iiit.ac.in,
chiranjeevi.yarra@iiit.ac.in

Abstract. Automatic syllable stress detection is helpful in assessing L2 learners' pronunciation. In this work, for stress detection, we propose a representation learning framework by jointly optimizing VAE and DNN. The obtained representations from the proposed VAE plus DNN framework are compared with the implicit representations learned from DNN-based stress detection. Further, we compare the obtained representations from VAE plus DNN with those obtained from autoencoder (AE) plus DNN, and sparse-autoencoder (SAE) plus DNN considering with/without implicit representations from DNN. We perform the experiments on the ISLE corpus consisting of English utterances from German and Italian native speakers. We observe that the detection performance with the learned representations from VAE plus DNN is significantly better than that with the state-of-the-art method without any representation learning with the highest improvement of 2.2%, 5.1%, and 1.4% under matched, combined, and cross scenarios, respectively.

Keywords: Syllable stress detection · Joint representation learning · Computer-assisted language learning

1 Introduction

The technological advancements showed their impact on teaching with the development of different computer-assisted language learning (CALL) based modules [3,14]. In recent years, applications related to CALL have shown benefits for second language learning. The reasons for the benefits include, 1) flexibility in its availability, 2) low cost of usage, and 3) ability to provide personalized learning [24]. However, developing robust modules for CALL is a challenging task, mainly due to the variabilities in L2 learners' native languages and accents. One of the many different aspects that the CALL applications have been focusing on is the detection and diagnosis of prosodic errors such as stress/prominence and intonation [9,16,31] errors made by the L2 learners in their pronunciation. Syllable

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 322–334, 2023.
https://doi.org/10.1007/978-3-031-48312-7_26

stress plays a critical role in communication to convey the meaning and intent of the message. Also, correct syllable stress placements in a word convey correct pronunciation. In this work, we consider the problem of automatic syllable stress detection which could be useful for downstream tasks such as CALL systems.

Syllable stress is referred to as the emphasis on a particular syllable in a word. Stressed syllables appear more prominent than unstressed syllables. In [1,5,27], it is stated that stress is mirrored through the changes in intensity, pitch, and duration. In [7], it is defined that the stressed syllable can be longer in relative duration and with greater physical intensity than the unstressed syllables but pitch movement does not always contribute to stress. Also in the literature, there is no strong agreement on the definition of stress in terms of acoustics for non-native English learners. Aoyama et al. [1] hypothesized that Japanese speakers rely more on differences in F0 compare to intensity and duration to indicate stress in English. Because of the native language influence, the production and perception of L2 will differ which in turn affects the acoustic parameters responsible for stress perception [15]. It highlights the need for a clearer and more consistent representation of stress.

Typically, automatic syllable stress detection has a feature extraction step followed by a machine learning (ML) based classification step. In the literature, various methods were proposed for better performance at both steps. At the model level, different ML algorithms like support vector machines (SVM), deep neural networks (DNN), convolutional neural networks (CNN), and attention networks were used for stress detection. Johnson et al. [10] used five different machine learning classifiers namely, neural networks, SVM, decision tree, bagging, and boosting algorithms for automatic detection of Brazil's prominence syllables with seven sets of different acoustic features embedding with variations in intensity, pitch, and duration. Arnold et al. [2] used random forests for prominence detection in the German language. Tian et al. [30] used an attention-based neural network and bidirectional LSTMs for stress detection problem using Mel frequency cepstral coefficients (MFCCs), energy, and pitch features. Ruan et al. [22] performed stress detection using a transformer network. Further, there were attempts on non-native speech in French, Spanish, and Mandarin with SVM for stress detection using acoustic and context-based features [4,8]. The above works are based on different models and features. Also, there are works [27,31] that focused only on feature level to extract the best features that can capture the syllable prominence.

Neural networks are often seen as a black box and it is difficult to interpret the kind of representation that the network is learning. Neural network architectures like DNN, CNN, and LSTM involves complex, nonlinear, and structured dependencies and they have been gaining popularity in different speech applications with their better performance over traditional ML methods. In automatic classification tasks, these neural networks learn representations implicitly, referred to as implicit representations, of the given input and map it to a specific class by learning weights. Even though these networks are known for learning task-specific implicit representations, they are sensitive to factors like variation and entanglement [6] in the data which can't be eliminated in real data scenarios.

One of the ways to overcome these factors is by learning a good representation of the data explicitly, prior to the classification task. To resolve this, AEs [23] were proposed and used in several applications for learning representations explicitly [28], referred to as explicit representations. But, AEs are not consistent in generating disentangled representation and regularized latent space. By addressing these issues, variational autoencoders (VAE) [13] gained attention in the field of computer vision [17] and speech processing applications [18,25] to learn the disentangled explicit representation of the data.

Obtaining combined representations by incorporating both the explicit and implicit representations through a single framework would benefit the stress detection task. However, to the best of our knowledge, there is no work that learns the combined as well as the explicit representations from the acoustic features for the stress detection task. To obtain combined (explicit and implicit) representations, we propose to optimize VAE and DNN in a joint learning framework.

In this work, we analyze the representations in a task-specific manner using acoustic and context-based features by modelling in two ways. First, we consider a DNN which implicitly learns representation from state-of-the-art acoustic and context-based features and performs classification. Second, we use the proposed representation learning framework jointly with VAE and DNN to obtain effective explicit and implicit representations for the stress detection task. Further, we analyze the effectiveness of the jointly learned representations obtained with VAEs compare to those obtained with other autoencoders namely, simple autoencoder (AE), and sparse autoencoder (SAE). We perform experiments on the ISLE corpus which consists of polysyllabic English words uttered by nonnative speakers of German, and Italian. We conduct the experiments in three scenarios: 1) matched: train and test data are from the same language, 2) combined: train data is from both the speakers' but tested on each of them separately, and 3) cross: trained with German and tested on Italian, and vice-versa. The jointly learned representations from VAE outperform the state-of-the-art method (without any representation learning) and implicit representations from DNN for stress detection. We found an absolute improvement in the classification accuracies by 2.2% & 1.2%, 5.1% & 1.1%, and 1.4% & 1.1% on German & Italian under matched, combined, and cross scenarios, respectively.

Table 1. Details of train and test splits of GER and ITA showing the number of stressed and unstressed syllables.

	Train		Test	
	#Stressed	#Unstressed	#Stressed	#Unstressed
GER	3076	3905	2756	3492
ITA	4408	5854	2148	2754

2 Database

For the experiments in this work, we consider ISLE [20] corpus. From this corpus, we consider 7834 speech utterances from 46 non-native speakers learning English where each speaker uttered 160 sentences following the work [31]. Out of 46 speakers, 23 are German (GER), and 23 are Italian (ITA). The entire audio was phonetically annotated by a team of five linguists to reflect the speakers' pronunciation. Using the automatic force alignment process, each utterance is phonetically aligned. We use P2TK [26], syllabification software to obtain syllable transcriptions from the phone transcriptions. From the syllable transcriptions, we obtain the aligned syllable boundaries using aligned phone boundaries. Syllable stress markings were also manually labeled while ensuring only one stressed syllable for each word. Labelling resulted in a total of 48868 syllables as stressed and 16693 syllables as unstressed. For the experiments, we consider data containing all polysyllabic words which result 12388 stressed and 16005 unstressed syllables. Train and test splits of both GER and ITA are done by balancing the speakers' nativity, age, sex, and proficiency [20]. Table 1 shows the details of the train test splits considered in the experiments.

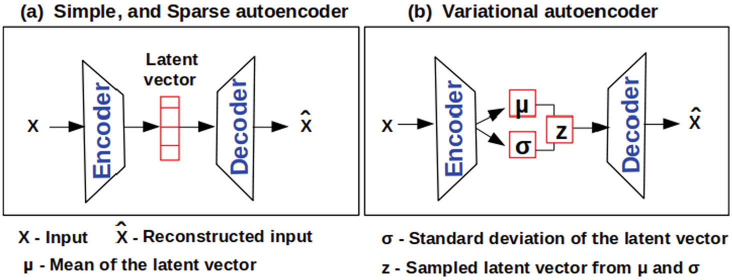
(a) Simple, and Sparse autoencoder
(b) Variational autoencoder
X
Encoder
Latent vector
Decoder
X̂
X
Encoder
μ
σ
z
Decoder
X̂
X - Input
X̂ - Reconstructed input
μ - Mean of the latent vector
σ - Standard deviation of the latent vector
z - Sampled latent vector from μ and σ

Fig. 1. Block diagrams of AE, SAE, and VAE.

3 Methodology

In this work, we consider VAE to learn the representations from the input features for the stress detection task. The VAEs are part of the autoencoder family, which includes AE, and SAE. In this section, we first briefly review AE, SAE, and VAE networks and then describe the framework of joint learning with VAE and DNN for syllable stress detection task.

3.1 Simple Autoencoder (AE):

Figure 1(a) illustrates the basic architecture of a simple autoencoder. It consists of an encoder and a decoder. The encoder encodes the d dimensional input feature vector X into a low dimensional latent vector and the decoder decodes the corresponding feature vector $\hat{X}$ from the latent vector. The entire encoder-decoder architecture is trained on the loss function which encourages the model to reconstruct the input from the latent vector at the output. Equation 1 shows the AE loss function, which is the mean square error between the encoder input and decoder output.

$$AE\,loss = (X - \hat{X})^2 \tag{1}$$

3.2 Sparse Autoencoder (SAE):

The autoencoders are usually prone to noise and learn more redundant information. In order to overcome this, sparse autoencoders were proposed. The sparse autoencoder architecture is the same as the autoencoder in Fig. 1(a) which has an encoder and decoder network. However, the loss function varies from AEs to SAEs. The SAE loss function includes a regularizer besides MSE loss in AE for penalizing the redundant information learning. The regularizer penalizes unnecessary nodes and activates selective nodes in the hidden layers of the encoder and decoder to avoid learning redundant information. Equation 2 shows the SAE loss function, in which the regularizer cost can be L_p norm (p=1 or 2) or Kullback-Leibler ($\mathbb{KL}$) - divergence on the parameters of encoder and decoder networks.

$$SAE\,loss = (X - \hat{X})^2 + regularizer \tag{2}$$

3.3 Variational Autoencoder (VAE):

Figure 1(b) illustrates the architecture of VAE. In VAE, for a given input vector X, unlike a fixed latent vector in AE and SAE encoder, $q_\theta(z|X)$ encodes the input feature vector to a latent vector space with a predefined random distribution ($p(z)$), typically a Gaussian density function with the mean μ and standard deviation σ. The decoder has two steps, the first step randomly samples the latent vectors $\mathbf{z}$ from the encoded latent space distribution using a reparametrization trick that uses unit normal Gaussian distribution, $z = \mu + \epsilon \cdot \sigma$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The second step decodes the input feature vector $\hat{X}$ from the latent vector $\mathbf{z}$. Equation 3 shows the loss function for VAE, which is defined considering two objectives. 1) Reconstructing the input (MSE), 2) Constraining the latent space to Gaussian distribution with KL-divergence. With this formulation, VAEs have shown great success in the field of computer vision [11]. Further, these have gained attention in speech processing analysis requiring latent representation learning. Thus, we believe the learned representations from the VAE could be robust for the stress detection task.

$$VAE\,loss = (X - \hat{X})^2 + \mathbb{KL}(q_\theta(z|X)||p(z)) \tag{3}$$

3.4 Joint Learning with VAE and DNN

Figure 2 illustrates the block diagram of the joint learning with VAE and DNN for syllable stress detection[1]. It has two flows, the first one is training flow

[1] https://github.com/Prasanth-Sai-Boyina/Syllable_stress_detection.

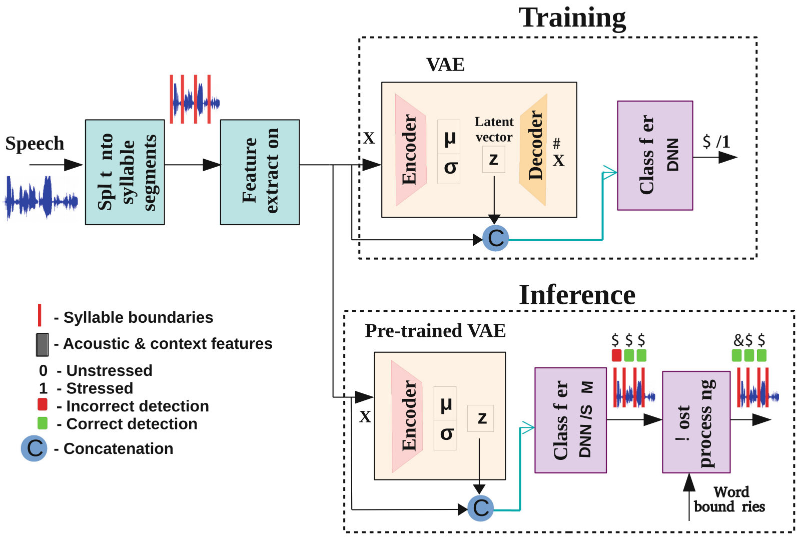

Fig. 2. Block diagram of joint learning approach.

and the second one is testing/inference flow. The two steps associated with the first two blocks are common for both training and testing flows. The first step obtains the syllable segments for a given speech utterance considering respective syllable transcriptions and their aligned boundaries. The second step computes input features for both training and testing. During training, we feed the input features to the VAE to learn the latent representations. The representations are learned by jointly training the VAE and the DNN classifier, which take latent representation and input features together (shown in the figure with blue lines) as input and stress markings (stressed and unstressed) as the output.

This joint training distinguishes our work from the typical training considered in the VAE. Equation 4 shows the loss function for the joint learning approach, which is defined considering two terms. 1) VAE loss consisting of MSE and KL-divergence, 2) Cross entropy (CE) loss between the predicted label ($\hat{Y}$) and ground truth (Y). λ_1 and λ_2 are the weight parameters. We hypothesize that by jointly optimizing the loss functions of VAE and DNN, we can learn the task-specific representations that would be robust for the detection task.

$$Joint\ loss = \lambda_1(VAE\ loss) + \lambda_2(CE(Y, \hat{Y})) \tag{4}$$

These representations consist of 1) representations that are explicit to DNN i.e. the latent representations learned by VAE, and 2) representations that are implicitly learned by DNN. Thus, we consider the proposed approach uses both explicit and implicit representations for the detection task. On the other hand, the learned latent representations are considered as explicit to the DNN when VAE and DNN are jointly trained without the concatenation step shown in Fig. 2.

After training VAE and DNN jointly, we extract the latent representations for the test data from the trained encoder, as shown in the inference block of the figure. We then use latent representation along with the input features as input (or only latent representation) to detect the syllable stress using DNN and SVM classifiers separately. The detected stress markings are post-processed to ensure that each polysyllabic word has only one stressed syllable following the work proposed by Yarra et al. [32].

4 Experimental Setup

In this study, both GER and ITA speakers' data is split into two non-overlapping sets namely, train and test sets. For the train set, following the previous work [31], we consider 1st-12th & 1st-13th speakers data and 13th-23rd & 14th-23rd speakers data for test set respectively for GER, and ITA [9]. Table 1 presents the details of syllable count in train and test conditions for both GER and ITA. We consider the state-of-the-art 19-dim acoustic-based features along with 19-dim binary features representing context dependencies following the work by Yarra et al. [32]. We consider their method, which uses an SVM classifier in the detection task as the baseline. We perform experiments in a 5-fold cross-validation setup where the train set is equally split into five groups, and the number of stressed and unstressed syllables are similar across five groups. In each fold, we use four sets for training, and one set for validation following a round-robin fashion. We normalize the training and testing set with the mean and standard deviation of the vectors obtained from the training set.

Architecture Details: The approach that we consider for representation learning jointly with VAE and DNN and classification with DNN/SVM is referred to as VAE+classifer (x + y; x represents the autoencoder used for learning task-specific representations jointly with DNN, y represents the classifier in the test time, either DNN or SVM that is used for classification with the learned representations from x). In the proposed approach, along with the VAE, we analyse the latent representations learned with simple AE, and SAE jointly with DNN and the corresponding networks are referred to as AE+classifer, and SAE+classifer, respectively. The DNN model in each of these consists of 8 hidden layers. We consider *Relu* [21] as activation function for the hidden layers and *Adam* [12] as optimizer. Binary cross-entropy is the loss function in DNN. AE and SAE consists of 2 hidden layers in encoder and decoder with *Relu* activation function. In SAE, we use L1 regularizer in one of the hidden layers of encoder. VAE consists of 1 hidden layer each in encoder and decoder with *Relu* activation function. All the DNN, VAE, AE, and SAE parameters like number of layers, and number of nodes in each layer are optimal and we choose them by maximizing the performance on the validation set. The optimal values of the parameters λ_1 and λ_2 in joint loss are found to be 0.53 and 0.47, respectively. For the SVM, we consider radial basis kernel and the parameter C by optimizing on the validation set. We consider an average of classification accuracies on the test set obtained from five training folds as a performance metric. We perform experiments in three different scenarios. 1) **matched:** We train with GER & ITA train sets and test on the GER & ITA test sets, respectively, 2) **combined:** We train with pooled data of GER and ITA, and test on GER and ITA test sets separately, and 3) **cross:** We train with GER(ITA) train set and test on ITA(GER) test set.

5 Results and Discussion

We analyze the learned representations – 1) both explicit and implicit, 2) implicit, 3) explicit, with the accuracies shown in Table 2 and Fig. 4. Table 2 reports the average classification accuracies with (in brackets) and without postprocessing obtained from baseline, DNN, VAE+DNN, and VAE+SVM on GER and ITA with acoustic (A) and acoustic plus context features (A+C) under all three scenarios. The results with VAE+DNN indicates the effectiveness of explicit and implicit representations combination. The results with DNN indicate the effectiveness of implicit representations. The explicit representations are analyzed with Fig. 4 by computing the accuracies without performing concatenation in Fig. 2 during testing/inference.

5.1 With Explicit and Implicit and Implicit Representations

Under Matched Scenario: From Table 2 under matched scenario, it is observed that in all the cases the accuracies obtained from VAE+DNN higher than those from baseline, DNN, and VAE+SVM with and without postprocessing. The highest improvements are found to be 2.2% & 1.2% and 1.9% & 1.4% on GER & ITA considering acoustic and acoustic plus context features, respectively. This indicates the benefit of both explicit and implicit representation compared to baseline (without any representations) and DNN (only with implicit representations). Further, the higher accuracies with DNN over baseline indicate the benefit of implicit representations. The higher accuracies with VAE+DNN compared to VAE+SVM indicate the effectiveness of implicit representations from DNN over SVM. The higher accuracies with the acoustic plus context features compared to acoustic features with the representation learning approach are consistent with the findings from the literature [27,29,32]. Altogether supports the benefit of the representation learning in stress detection task.

Under Combined Scenario: The comparisons made under matched scenario across baseline, DNN, VAE+DNN, and VAE+SVM are consistent under combined scenario also. From the table, the highest accuracies in GER and ITA are found in the combined scenario and those are 94.1% and 94.2%, respectively, obtained from VAE+DNN considering acoustic plus context features with postprocessing. Further, while comparing the accuracies between matched and combined scenarios, the accuracies are higher under combined than those under matched with VAE+DNN and DNN but not in all cases of baseline and VAE+SVM. Both these together suggest that the combined scenario has an advantage for the stress detection task compared to the matched scenario and

Table 2. Classification accuracies with (without) postprocessing considering acoustic (A) and acoustic plus context (A+C) features under three different scenarios.

Test data	Train scenario	**SVM**		**DNN**		**VAE+DNN**		**VAE+SVM**	
		A	**A+C**	**A**	**A+C**	**A**	**A+C**	**A**	**A+C**
GER	**Matched**	83.5 (80.3)	92.3 (88.7)	84.6 (81.1)	92.6 (88.3)	**85.7** **(82.1)**	**93.5** **(90.4)**	84.5 (81.3)	92.4 (89.1)
	Combined	83.2 (80.5)	89 (85.1)	84.1 (81.1)	92.7 (89.1)	**85.4** **(82)**	**94.1** **(90.9)**	84.6 (81.6)	92.3 (89.5)
	Cross	80.5 (77.7)	88.2 (84.5)	80.2 (77.4)	88.3 (84.5)	**80.9** **(78.1)**	**89.6** **(85.6)**	80.7 (78.5)	87.5 (83.6)
ITA	**Matched**	82.7 (80.5)	91.5 (88.2)	82.8 (81.1)	91.7 (87.7)	**84.6** **(82.3)**	**92.9** **(89.5)**	83.7 (81.4)	91.3 (88.4)
	Combined	83.4 (81)	93 (89.8)	83.8 (81.6)	93.3 (89.2)	**85.4** **(82.6)**	**94.2** **(90.6)**	84.5 (82.2)	92.6 (89.7)
	Cross	82.1 (79.3)	90.7 (86.6)	82.7 (79.4)	90.9 (86.2)	**83.6** **(80)**	**91.8** **(86.8)**	81.9 (79)	87.6 (84)

shows that VAE+DNN and DNN utilize the extra data in the stress detection task whereas baseline and VAE+SVM failed to do so.

Under Cross Scenario: From Table 2, it is observed that there is a drop in accuracies under cross scenario compared to those under matched scenario in baseline, DNN, VAE+DNN, and VAE+SVM. This could be due to the mismatch in the nativity. But the VAE+DNN is performing better over the baseline, and DNN in GER, and ITA in all the cases considering both with and without postprocessing. This indicates that the explicit and implicit representations learned with VAE+DNN could be independent of speakers' nativity and effective in learning the stress detection task-specific cues through the representations. From all the above comparisons, the significant improvements with the VAE+DNN over baseline, DNN, and VAE+SVM among all three scenarios indicate the robustness of the explicit and implicit representations for stress detection.

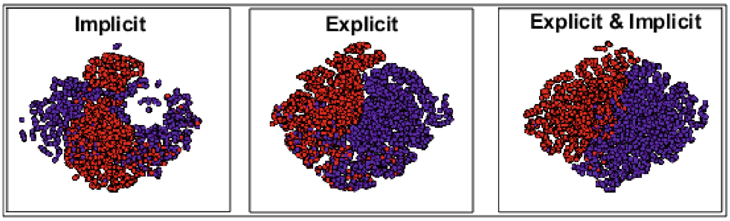

Fig. 3. t-SNE visualizations of learned representations under three approaches. ● Class 0, ● Class 1.

5.2 With Explicit Representations

Representations can be learned through different types of autoencoders. In this work, we consider VAE due to its effectiveness in learning representations. In order to analyze the same, we also compute the accuracies with the representations learned from other types – AE and SAE. We perform the analysis considering only explicit representations (without concatenation in Fig. 2) and comparing them with DNN and the baseline. The accuracies obtained from the autoencoders' (AE, SAE, and VAE) explicit representations and those from DNN and the baseline have similar trend across both GER and ITA, so for better readability, we present the accuracies averaged across GER and ITA.

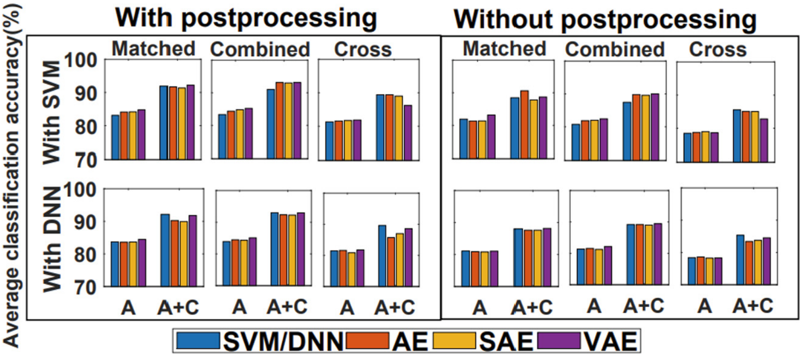

Fig. 4. Comparison of average classification accuracies obtained from explicit representations learned with AE, SAE, and VAE using classifier as SVM (first row) and DNN (second row) separately.

Figure 4 presents the average classification accuracies considering acoustic, and acoustic plus context features under all three scenarios with and without postprocessing. Each bar height represents average classification accuracy. The first and second rows correspond to the classification accuracies considering the test classifier as SVM, and DNN, respectively. From the figure, we observe that acoustic plus context features are significantly better than acoustic features with, and without postprocessing. From the first row, where the classifier is SVM, it is observed that classification with representation learning approaches (AE, SAE, and VAE) are higher than the baseline in majority of the cases. And there is an increasing trend in the performance among AE+SVM, SAE+SVM, and VAE+SVM in 3 out of 4 cases except in the cross scenarios. This indicates that the representations learned from VAE are comparable to and better than the other autoencoder types. On the other hand, a similar trend among the autoencoders is not consistent in the second row, where the classifier is DNN. Further, the accuracies with the DNN are higher than those with the AE+DNN, SAE+DNN, and VAE+DNN. This suggests that the explicit representations alone could be less effective compared to the implicit representations learned by

DNN. However, comparing Table 2 and Fig. 4, it is observed that the accuracies with the VAE+DNN considering explicit and implicit representation are higher than those with the DNN. Further, we observe that the accuracies with the VAE+DNN are higher than those with the AE+DNN and SAE+DNN considering explicit and implicit representations. These together indicate the benefit of the representations learned from VAE in the stress detection task considering the proposed explicit and implicit representation-based approaches compare to implicit, and explicit alone representations based approaches. The t-SNE [19] visualizations shown in Fig. 3 suggest that the explicit and implicit based representation learning approach is capable of discriminating the classes better.

6 Conclusion

In this work, we have considered a representation learning approach jointly with VAE and DNN for automatic syllable stress detection task using acoustic and context-based features. The learned representations include three sets of representations namely, 1) implicit, 2) explicit, and 3) explicit and implicit. The proposed joint learning approach learns both explicit and implicit representations. Experiments with ISLE corpus showed that stress detection performance with the proposed joint representation learning approach consistently performs better than the baseline, and DNN (implicit) in both GER and ITA under matched, combined, and cross-native scenarios. Further, representations learned from VAE were found to be better than those of AE, and SAE. In the future, we would like to investigate end-to-end based representation learning and self-supervised based representations for syllable stress detection to overcome the difficulty in manual labeling of stress markings.

References

1. Aoyama, K., Guion, S.G.: Prosody in second language acquisition. Language Experience in Second Language Speech Learning: in honor of James Emil Flege. Amsterdam (2007)
2. Arnold, D., Wagner, P., Baayen, R.H.: Using generalized additive models and random forests to model prosodic prominence in German. In: INTERSPEECH, Lyon, France, pp. 272–276. International Speech Communications Association (2017)
3. Bernhard, V., Schwab, S., Goldman, J.P.: Acoustic Stress Detection in Isolated English Words for Computer-Assisted Pronunciation Training. In: Proceedings of Interspeech 2022, pp. 3143–3147 (2022). https://doi.org/10.21437/Interspeech 2022–197
4. Christodoulides, G., Avanzi, M.: An evaluation of machine learning methods for prominence detection in French. In: INTERSPEECH, pp. 116–119 (2014)
5. Couper-Kuhlen, E.: An introduction to english prosody. (No Title) (1986)
6. Cunningham, P., Carney, J., Jacob, S.: Stability problems with artificial neural networks and the ensemble solution. Artif. Intell. Med. **20**(3), 217–225 (2000)
7. Cutler, A., Isard, S.D.: The production of prosody (1980)

8. Evin, D., Cossio Mercado, C., Torres, H.M., Gurlekian, J., Mixdorff, H.: Automatic prominence detection in Argentinian Spanish, Proceedings of Speech Prosody, Poznan, Poland, pp. 680–684 (2018)
9. Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., Precoda, K.: Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. Speech Commun. **69**, 31–45 (2015)
10. Johnson, D.O., Kang, O.: Automatic prominent syllable detection with machine learning classifiers. Int. J. Speech Technol. **18**(4), 583–592 (2015). https://doi.org/10.1007/s10772-015-9299-z
11. Kim, J.H., Zhang, Y., Han, K., Wen, Z., Choi, M., Liu, Z.: Representation learning of resting state fMRI with variational autoencoder. Neuroimage **241**, 118423 (2021)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Lewis, C.: The Role of Lexical Stress in English as a Lingua Franca in Southeast Asia. In: Pronunciation in Second Language Learning and Teaching Proceedings, vol. 12(1) (2022)
15. Li, A., Post, B.: L2 acquisition of prosodic properties of speech rhythm: evidence from l1 mandarin and German learners of English. Stud. Second. Lang. Acquis. **36**(2), 223–255 (2014)
16. Li, K., Mao, S., Li, X., Wu, Z., Meng, H.: Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Commun. **96**, 28–36 (2018)
17. Lin, C.C., Hung, Y., Feris, R., He, L.: Video instance segmentation tracking with a modified vae architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13147–13157 (2020)
18. Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N., Roberts, S.: Anomaly detection for time series using VAE-LSTM hybrid model. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4322–4326. IEEE (2020)
19. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
20. Menzel, W., et al.: The ISLE corpus of non-native spoken English. In: Proceedings of LREC: Language Resources and Evaluation Conference, vol. 2, pp. 957–964. European Language Resources Association (2000)
21. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
22. Ruan, Y., et al.: An end-to-end approach for lexical stress detection based on transformer. arXiv preprint arXiv:1911.04862 (2019)
23. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
24. Su, P.H., Wu, C.H., Lee, L.S.: A recursive dialogue game for personalized computer-aided pronunciation training. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 127–141 (2014)
25. Sun, G., et al.: Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6699–6703. IEEE (2020)
26. Tauberer, J.: P2tk automated syllabifier (2008)

27. Tepperman, J., Narayanan, S.: Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In: Proceedings (ICASSP) IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I-937. IEEE (2005)
28. Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069 (2018)
29. Umeda, N.: Vowel duration in American English. J. Acoustical Soc. Am. **58**(2), 434–445 (1975)
30. Xia, T., Rui, X., Huang, C.L., Chu, I.H., Wang, S., Han, M.: An Attention Based Deep Neural Network for Automatic Lexical Stress Detection. In: Global Conference on Signal and Information Processing (GlobalSIP), pp. 1–5. IEEE (2019)
31. Yarra, C., Deshmukh, O.D., Ghosh, P.K.: Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5845–5849. IEEE (2017)
32. Yarra, C., Ramanathi, M.K., Ghosh, P.K.: Comparison of automatic syllable stress detection quality with time-aligned boundaries and context dependencies. In: SLaTE, pp. 79–83 (2019)