STEPs-RL: Speech-Text Entanglement for Phonetically Sound Representation Learning

Prakamya Mishra
Independent Researcher
pkms.research@gmail.com

Abstract

In this paper, we present a novel multi-modal deep neural network architecture that uses speech and text entanglement for learning phonetically sound spoken-word representations. STEPs-RL is trained in a supervised manner to predict the phonetic sequence of a target spoken-word using its contextual spoken word's speech and text, such that the model encodes its meaningful latent representations. Unlike existing work, we have used text along with speech for auditory representation learning to capture semantical and syntactical information along with the acoustic and temporal information. The latent representations produced by our model were not only able to predict the target phonetic sequences with an accuracy of 89.47% but were also able to achieve competitive results to textual word representation models, Word2Vec & FastText (trained on textual transcripts), when evaluated on four widely used word similarity benchmark datasets. In addition, investigation of the generated vector space also demonstrated the capability of the proposed model to capture the phonetic structure of the spoken-words. To the best of our knowledge, none of the existing works use speech and text entanglement for learning spoken-word representation, which makes this work first of its kind.

1   Introduction

Speaking and listening are the most common ways in which humans convey and understand each other in daily conversations. Nowadays, the speech interface has also been widely integrated into many applications/devices like Siri, Google Assistant, and Alexa [14]. These applications use speech recognition-based approaches [12, 20, 4] to understand the spoken user queries. Like speech, the text is also a widely used medium in which people converse. Recent advances in language modeling and representation learning using deep learning approaches [28, 3, 8, 31] have proven to be very promising in understanding the actual meanings of the textual data, by capturing semantical, syntactical, and contextual relationships between the textual words in their corresponding learned fixed-size vector representations.

Such computational language modeling is difficult in the case of speech for spoken language understanding because unlike textual words, (1) spoken words can have different meanings of the same word when spoken in different tones/expressions [10], (2) it is difficult to identify sub-word units in speech because of the variable-length spacing and overlapping between the spoke-words [40], and (3) use of stress/emphasis on few syllables of a multi-syllabic word can increase the variability of speech production [32]. Although the textual word representations capture the semantical, syntactical, and contextual properties, they fail to capture the tone/expression. Using only speech/audio data for training spoken-word representations results in semantically and syntactically poor representations.

So in this paper, we propose a novel spoken-word representation learning approach called STEPs-RL that uses speech and text entanglement for learning phonetically sound spoken-word representations, which not only captures the acoustic and contextual features but also are semantically, syntacti-

arXiv:2011.11387v1 [cs.CL] 23 Nov 2020

cally, and phonetically sound. STEPs-RL is trained in a supervised manner such that the learned representations can capture the phonetic structure of the spoken-words along with their inter-word semantic, syntactic, and contextual relationships. We validated the proposed model by (1) evaluating semantical and syntactical relationships between the learned spoken-word representations on four widely used word similarity benchmark datasets, and comparing its performance with the textual word representations learned by Word2Vec & FastTexT (obtained using transcriptions), and (2) investigating the phonetical soundness of the generated vector space.

2   Related Work

Earlier, speech processing was done using feature learning-based models like deep neural networks (DNN) [33]. The DNN models were able to capture contextual and temporal information from the speech-based data after the introduction of sequential neural networks like RNNs [27, 17, 34], LSTMs [23, 29], Bi-LSTMs [42, 11], and GRUs [39, 35].

Recent research by Liu et al. [26] has presented the use of a transformer-based self-supervised speech representation learning approach called TERA that uses multi-target auxiliary tasks. TERA is trained by generating acoustic frame reconstructions; Schneider et al. [36] introduced wav2vec which is a CNN based model pre-trained in a unsupervised manner using contrastive loss to learn raw audio representations; Khurana et al. [22] explored the use of black-box variational inference for linguistic representation learning of speech using an unsupervised generative model; van den Oord et al. [30] proposed Contrastive Predictive Coding (CPC) for extracting representations from high dimension data by predicting future in latent space, using autoregressive models; Amodei et al. [1] presented an end-to-end deep learning model to recognize speech in two vastly different languages (English & Mandarin Chinese); Hsu et al. [19] proposed a novel variational autoencoder based model that learns disentangled and interpretable latent representations of sequential data in an unsupervised manner; Ling et al. [25] used BERT encoder for learning phonetically aware contextual speech representation vectors; Chen et al. [5] proposed a Word2Vec type sequence-to-sequence autoencoder model for embedding variable-length audio segments. Other works on learning fixed-length spoken-word vector representations that use multi-task learning include [6], [21], [7], [24] and [38].

3   Model

In this paper, we propose STEPs-RL: Speech-Text Entanglement for Phonetically Sound Representation Learning. STEPs-RL is a novel spoken-word representation learning approach which entangles speech and text based contextual information for learning phonetically sound spoken-word representations. The model architecture is shown in Figure 1. Given a target spoken-word represented by S^t, its left and right contextual spoken-words represented by $S_{ctx}^l = \{S^i\}_{t-1-m}^{t-1}$ & $S_{ctx}^r = \{S^i\}_{t+1}^{t+1+m}$ respectively (m represents the context window size), along with the textual word embeddings of the corresponding spoken-words represented by $W_{ctx}^l = \{W^i\}_{t-1-m}^{t-1}$, W^t & $W_{ctx}^r = \{W^i\}_{t+1}^{t+1+m}$, the proposed model tries to learn a vector representation of the target spoken-word that not only captures the semantic-based, syntax-based and acoustic-based information but also captures the phonetic-based information.

Here, a single spoken-word $S^i \in \mathbb{R}^{n \times d_{mfcc}}$ consists of a sequence of acoustic features represented by d_{mfcc}-dimensional Mel-frequency Cepstral Coefficients (MFCCs); $W^i \in \mathbb{R}^{d_w}$ represents the d_w-dimensional pre-trained textual word embedding of the corresponding spoken-word. Each of the spoken-word is padded with silence, so that they all consists of a sequence of n acoustic features.

Our approach uses Bidirectional-LSTM [37] for capturing the contextual information. Bidirectional-LSTM (also known as Bi-LSTM), uses two LSTM [16] networks ($\overrightarrow{LSTM}, \overleftarrow{LSTM}$) to capture contextual information in opposite directions (forward and backward) of a sequence ($t_1, ..t_T$). The final hidden representations corresponding to the sequence tokens is generated by concatenating ($\oplus$) the hidden representations ($\overrightarrow{h_i}, \overleftarrow{h_i}$) generated by both the LSTM networks. So the final hidden representation of the ith token can be represented as shown in equation 2.

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(t_i, \overrightarrow{h_{i-1}}), \quad \overleftarrow{h_i} = \overleftarrow{LSTM}(t_i, \overleftarrow{h_{i+1}}) \tag{1}$$

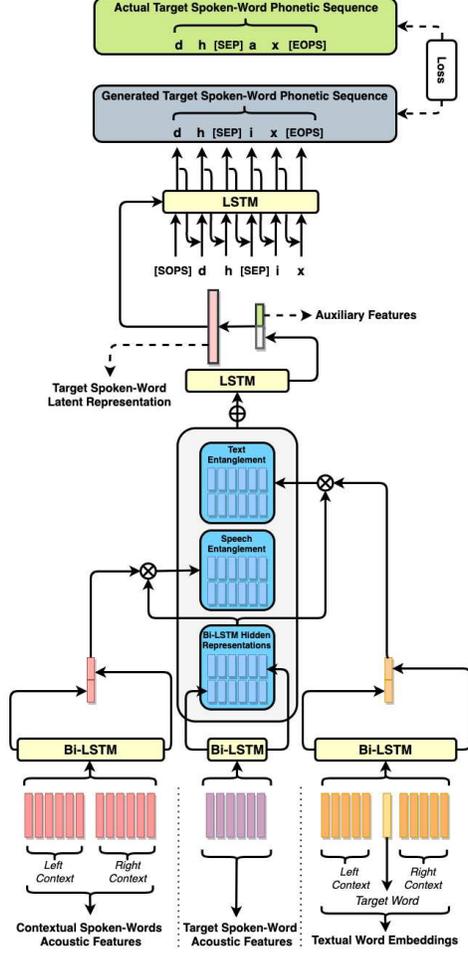$$h_i = \overrightarrow{h_i} \oplus \overleftarrow{h_i} \tag{2}$$

Figure 1: Illustration of the STEPs-RL model architecture.

STEPs-RL consist of three independent Bi-LSTM networks represented by $BiLSTM_C$, $BiLSTM_T$ and $BiLSTM_W$ to capture contextual information respectively from (1) The acoustic features of the left and right contextual spoken-words represented by S^l_{ctx} & S^r_{ctx}, (2) The acoustic features of the target spoken-word represented by S^t, and (3) The pre-trained textual word embeddings of the corresponding target spoken-word, left contextual spoken-words and right contextual spoken-words represented by W^t, W^l_{ctx} & W^r_{ctx} respectively.

$$h^C, \overrightarrow{o^C}, \overleftarrow{o^C} = BiLSTM_C([S^l_{ctx}, S^r_{ctx}]) \tag{3}$$

$$h^T, \overrightarrow{o^T}, \overleftarrow{o^T} = BiLSTM_T([S^t]) \tag{4}$$

$$h^W, \overrightarrow{o^W}, \overleftarrow{o^W} = BiLSTM_W([W^l_{ctx}, W^t, W^r_{ctx}]) \tag{5}$$

As shown in equations 3, 4 & 5, all the three Bi-LSTM networks generate a final hidden state representation corresponding to each timestamp (h^C, h^T, h^W), a final output of the corresponding forward LSTM network ($\overrightarrow{o^C}$, $\overrightarrow{o^T}$, $\overrightarrow{o^W}$), and a final output of the corresponding backward LSTM network ($\overleftarrow{o^C}$, $\overleftarrow{o^T}$, $\overleftarrow{o^W}$). The final forward and backward outputs of $BiLSTM_C$ & $BiLSTM_W$ are concatenated to generate f^C & f^W respectively, which will later act as context vectors during the entanglement of speech and text.

$$f^C = \overrightarrow{o^C} \oplus \overleftarrow{o^C}, \quad f^W = \overrightarrow{o^W} \oplus \overleftarrow{o^W} \tag{6}$$
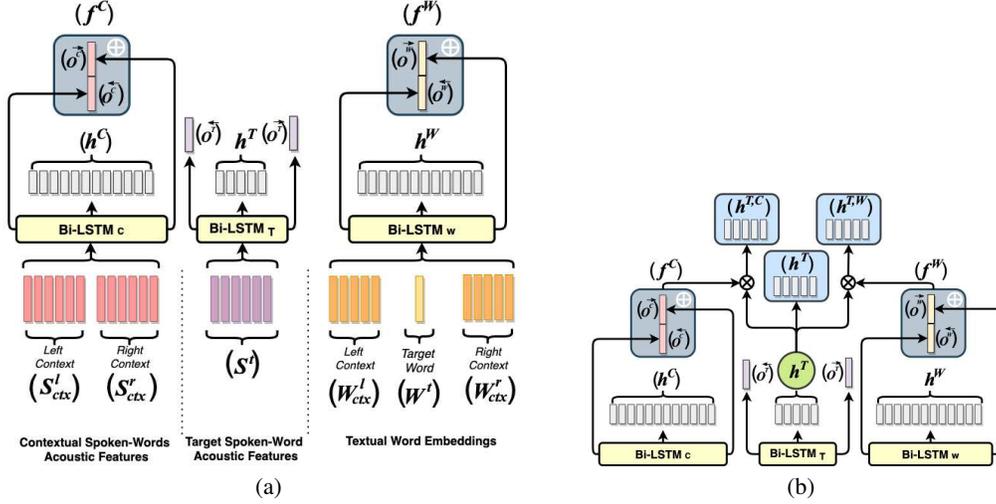
Figure 2: (a) STEPs-RL Phase 1: Each of the individual Bi-LSTM captures contextual information. (b) STEPs-RL Phase 2: Speech & Text entanglement with target spoken word.

For intuition (as shown in Figure 2a), f^C represents the final contextual representation of the spoken-words present in context of the target spoken-word, and f^W represents the final semantical and syntactical contextual representation of all the corresponding textual words. In other words, f^C captures the acoustic/speech-based contextual information whereas f^W captures the text-based contextual information. Both f^C & f^W, are then used to entangle speech and text-based contextual information with the target spoken-word by generating new speech and text entangled bidirectional hidden state representations ($h^{T,C}$ & $h^{T,W}$) of the target spoken-word using the hidden representations generated by $BiLSTM_T$, as shown in equations 7, 8, 9 & 10.

$$h^{T,C} = [h_1^{T,C}, h_2^{T,C}, ..., h_n^{T,C}] = h^T \otimes f^C \tag{7}$$

$$h_i^{T,C} = \alpha_i^{T,C} \times h_i^T \tag{8}$$

$$h^{T,W} = [h_1^{T,W}, h_2^{T,W}, ..., h_n^{T,W}] = h^T \otimes f^W \tag{9}$$

$$h_i^{T,W} = \alpha_i^{T,W} \times h_i^T \tag{10}$$

In the above equations, ($\otimes$) represents an element wise attention function; $h^{T,C}$ & $h^{T,W}$ represents the newly generated speech-entangled and text-entangled hidden representations respectively; $\alpha_i^{T,C}$ & $\alpha_i^{T,W}$ represents the speech-entangled and text-entangled attention scores respectively, corresponding to the ith timestamp of the hidden representations generated by $BiLSTM_T$. The attention scores $\alpha_i^{T,C}$ & $\alpha_i^{T,W}$ are generated by taking the dot product ($\bullet$) of each of the timestamps of h^T with the context vectors f^C & f^W respectively, as shown in equation 11. Same is illustrated in Figure 2b.

$$\alpha_i^{T,C} = h_i^T \bullet f^C, \quad \alpha_i^{T,W} = h_i^T \bullet f^W \tag{11}$$

Next, the proposed model uses the newly generated speech-entangled and text-entangled hidden representations $h^{T,C}$ & $h^{T,W}$, along with the original bidirectional hidden state representations h^T of the target spoken-word (generated from $BiLSTM_T$), to generate a latent vector representation z of the target spoken-word by stacking (illustrated in Figure 3) all these three hidden representations on top of each other and passing it through a simple encoder LSTM network $\overrightarrow{LSTM_{encode}}$.

$$z = \overrightarrow{LSTM_{encode}}([h^{T,C} \oplus h^{T,W} \oplus h^T]) \tag{12}$$

$$z_{new} = zW_1 + z_{aux}W_2 + B \tag{13}$$

In equation 12, z represents a fixed size latent vector which is the output of the encoder LSTM network. To add more information about the speaker, the proposed model linearly combines the
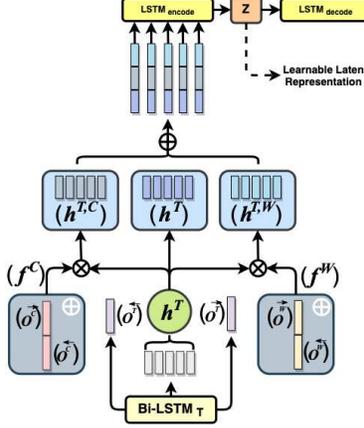
Figure 3: STEPs-RL Phase 3: Latent representation learning

latent vector with an auxiliary vector z_{aux} to generate a new latent representation z_{new} of the target spoken-word. This new latent representation $z_{new} \in \mathbb{R}^{d_e}$ is a d_e-dimensional vector representation that the proposed model tries to learn. In equation 13, $W_1 \in \mathbb{R}^{d \times d_e}$ and $W_2 \in \mathbb{R}^{d_a \times d_e}$ represents the combination weights and B represents the bias. These weights and biases are learnable in nature. The auxiliary vector $z_{aux} \in \mathbb{R}^{d_a}$ is a one-hot vector of size d_a that consists of information related to the speaker's gender/dialect or both. Such an auxiliary vector was introduced because usually, the pronunciation of different words usually depends on the speaker's gender and dialect and hence can help learn phonetically sound spoken-word representations.

Next, the proposed model uses a decoder LSTM network $\overrightarrow{LSTM_{decode}}$ to predict the sequence of phonetic symbols $Y=([y_1, ..., y_k])$ of the corresponding target spoken-word using the above generated latent representation of the target spoken-word z_{new}, as shown in equation 14 & 15.

$$P_\theta(y_i|Y_{<i}, z_{new}) = \Upsilon(h_i^d, y_{i-1}) \tag{14}$$

$$h_i^d = \Psi(h_{i-1}^d, y_{i-1}) \tag{15}$$

Here, Ψ represents a function that generates the hidden vectors h_i^d (hidden state representations of the decoder network), and Υ represents a function that computes the generative probability of the one-hot vector y_i (target phonemic symbol). The hidden vector h_i^d is z_{new}, and y_i is the one-hot vector of "[SOP]" when $i=0$. Here "[SOP]" represent the start of phoneme token. The proposed model uses cross-entropy as its training loss function as shown in equation 16, where cross-entropy loss L is computed using the actual target spoken-word phonetic sequence ($Y=([y_1, ..., y_k])$) and the predicted target spoken-word phonetic sequence ($\hat{Y}=([\hat{y_1}, ..., \hat{y_k}])$).

$$L(Y, \hat{Y}) = \sum_i^k y_i \log \frac{1}{\hat{y_i}} \tag{16}$$

4   Experimental Setup

4.1   Dataset

For our experiments, we used the DARPA TIMIT acoustic-phonetic Continuous Speech Corpus [9]. This corpus contains 16kHz audio recordings of 630 speakers of 8 major American English dialects of which approximately 70% were male and 30% were female, as shown in Table 1. The corpus consists of 6300 (5.4 hours) phonetically rich utterances by different speakers (10 by each speaker) along with their corresponding time-aligned orthographic, phonetic, and word transcriptions. All the recordings were segmented according to the spoken-word boundaries using the transcriptions and were paired with their left and right context spoken-words and their corresponding textual words along with the

Table 1: Gender and dialect distribution of the speakers in TIMIT speech corpus.

Gender	Dialect							
	1	2	3	4	5	6	7	8
Male	63%	70%	67%	69%	63%	65%	74%	67%
Female	27%	30%	27%	31%	37%	35%	26%	33%
Total	**8%**	**16%**	**16%**	**16%**	**16%**	**7%**	**16%**	**5%**

Table 2: Phonetic sequence prediction results on the TIMIT speech corpus. We present here the comparison of testing set accuracy (%) of the STEPs-RL model using different sets of auxiliary information (gender (**G**), dialect (**D**)) with the base STEPs-RL model using no auxiliary information. The comparison is done for different textual word embedding sizes d_w={50,100,300}, different spoken-word latent representation sizes d_e={50,100,300} and different word embeddings like Word2Vec (*w*) and FastText (*f*). The best performance in each configuration is marked in **bold**, row of the best performing model is highlighted in grey, the overall best performance is further marked in red and its configuration is marked in blue.

Spoken-Word Latent Rep. Size (d_e) →	$d_e = 50$						$d_e = 100$						$d_e = 300$					
Textual Word Embedding Size (d_w) →	$d_w = 50$		$d_w = 100$		$d_w = 300$		$d_w = 50$		$d_w = 100$		$d_w = 300$		$d_w = 50$		$d_w = 100$		$d_w = 300$	
Textual Word Embeddings Used →	*w*	*f*	*w*	*f*	*w*	*f*	*w*	*f*	*w*	*f*	*w*	*f*	*w*	*f*	*w*	*f*	*w*	*f*
STEPs-RL + No auxiliary information	71.67	73.24	75.22	78.41	84.76	84.98	73.44	76.72	78.36	81.23	86.90	86.92	75.22	79.75	80.59	83.73	87.31	86.90
STEPs-RL + D	83.87	84.01	86.11	86.36	85.89	87.89	86.83	86.85	86.75	87.54	87.23	88.00	81.50	82.05	85.90	86.36	87.40	87.99
STEPs-RL + G	87.93	86.97	87.44	85.60	87.16	88.28	87.15	87.12	87.98	**88.20**	**88.92**	**88.54**	87.16	87.30	88.91	88.45	88.10	88.59
STEPs-RL + D + G	**88.91**	**88.04**	**88.10**	**88.28**	**88.94**	**88.78**	**88.27**	**88.90**	**89.14**	88.08	88.59	89.18	**89.47**	**89.21**	**89.35**	**89.38**	**88.63**	**89.41**

phonetic sequence of the target spoken-word. All the spoken-word utterances were represented by their MFCC representations and the textual words were represented by their pre-trained textual word embeddings, where the MFCC representations and the textual word embeddings were of the same size (d_{mfcc}=d_w). One-hot encoded dialect (8-dimensional) and gender (2-dimensional) vectors were used as auxiliary information vectors.

4.2 Training Details

We used the standard train (462 speakers and 4956 utterances) and test (168 speakers and 1344 utterances) set of the TIMIT speech corpus for training and testing the proposed model. Due to computational resource limitations, a context window size of 3 was used. In all the experiments the MFCC representations and the textual word embeddings were of the same size (d_{mfcc}=$d_w \in$ {50,100,300}). For the textual word embeddings, the proposed model used two different widely used pre-trained word embeddings i.e, (1) Word2Vec [28], which are word-based embeddings, and (2) FastText [3], which are character-based embeddings. For all the experiments, the proposed model was trained for 20 epochs using a mini-batch size of 100. The initial learning rate was set to 0.01 and Adam optimizer was used for optimization. The Bi-LSTM and LSTM nodes were regularised using an L2 regularizer with a penalty of 0.01. Early stopping was used to avoid over-fitting. The size of the target spoken-word latent representation z_{new} was set to 50-, 100- & 300 for comparison. All the spoken-words were represented by a sequence of 50 phonetic symbols using the original unique 27 phonetic symbols present in the corpus along with our four newly introduced symbols ("**[SOPS]**" for the start of each phonetic sequence, "**[SEP]**" for separation/space between phonetic symbols, "**[PAD]**" for padding and "**[EOPS]**" for the end of each phonetic sequence).

5 Results

For evaluation, we first tested the proposed model on the phonetic sequence prediction task with different spoken-word latent representation & textual word embedding sizes, and also tested the performance of the model using different types of textual word embeddings (Word2Vec & FastText). We compared the phonetic sequence prediction accuracy (%) of the base STEPs-RL model (w/o any auxiliary information) with its variants that use different sets of auxiliary information like gender/dialect or both. The results are shown in Table 2. It was observed that increasing the spoken-word representation size resulted in better performance but was not so evident in the case of textual word embedding size. It was also observed that in general using Word2Vec textual word embeddings achieved better results compared to using FastText textual word embeddings. The

Table 3: (a) Examples of the phonetic sequences generated by STEPs-RL+D+G model. (b) Performance of STEPs-RL+D+G compared to Word2Vec & FastText on four benchmark word similarity datasets.

(a)

Ground Truth Word	Ground Truth Phonetic Sequence	Generated Phonetic Sequence
that	dh ae tcl	dh ax tc l
shelter	s sh eh l tcl t axr	s tsh ah el t er
near	n ih axr	n ehh axr
heck	hv eh kcl k	hv ah ncl k

(b)

Dataset	# Word Pairs	STEPs + D + G (ρ)	Word2Vec (ρ)	FastText (ρ)
SimLex	136	0.2552	0.2792	0.3375
MTurk	55	0.5312	0.6536	0.6771
WS	17	0.2726	0.3248	0.3082
Verb	26	0.3260	0.3508	0.3676

addition of auxiliary information like dialect and gender showed clear improvements in accuracy when compared to the base STEPs-RL model, validating the use of this type of auxiliary information for spoken-word representation learning. It was also found that STEPs-RL was able to perform best when it used both dialect (**D**) and gender (**G**) together in its auxiliary vector (**STEPs-RL+D+G**). So for further evaluations, we will only consider the target spoken-word representations generated from the STEPs-RL+D+G model using the configurations marked blue in Table 2. Table 3a illustrates examples of four different spoke-words along with their actual corresponding phonetic sequences and the phonetic sequences predicted by the STEPs-RL+D+G model. These examples demonstrate the ability of the STEPs-RL+D+G model to encode phonetic-based information in their corresponding latent representations.

To further evaluate the latent representations generated from STEPs-RL+D+G, we use intrinsic methods to test the semantic or syntactic relationships between these generated latent representations of the spoken-words present in the corpus. To do so, we use 4 benchmark word similarity datasets and compare the performance of the spoken-word representations generated from STEPs-RL+D+G with the representations generated by text-based language models (Word2Vec & FastText) trained on the textual transcripts. The word similarity datasets include SimeLex-999 [15], MTurk-771 [13], WS-353 [41] and Verb-143 [2]. These datasets contain pairs of English words and their corresponding human-annotated word similarity ratings. The word similarities between the spoken-words (in case of STEPs-RL+D+G) and the textual-words (in case of Word2Vec and FastText) were obtained by measuring the cosine similarities between their corresponding representation vectors.

Table 3b reports Spearman's rank correlation coefficient ρ between the human rankings and the ones generated by STEPs-RL+D+G, Word2Vec, and Fastext. Since there were many words present in these datasets which were not present in the TIMIT speech corpus, only those word pairs were considered in which both the word were present in the TIMIT speech corpus. Table 3b shows that the performance of the spoken-word representations generated from STEPs-RL+D+G was comparable to the performance of textual word representations generated from Word2Vec and FastText. This demonstrates that our proposed model was also able to capture semantic-based and syntax-based information, although the scores were slightly less compared to Word2Vec and FastText.

We believe that the primary reason for this difference is the disparity in the way different speakers speak. The same word can be spoken in different ways and can have different meanings based on the tone and expression which may in return lead to an entirely different representation for the same word. In addition to that, these word similarity datasets are for the textual words, which do not take into account the tone and the expression aspect. Also, to the best of our knowledge, no other such word similarity dataset exists for the spoken-words. So keeping in mind these issues, the performance of the proposed model validates its ability to capture semantical and syntactical information in the representations it generates.

Next, we try to investigate the phonetical soundness of the vector space generated by the proposed model. A vector space can be said to be phonetically sound if the spoken-word representations of the words having similar pronunciations are present close to each other in the vector space. For this investigation we use 2 sets of randomly chosen word pairs:

- **Set 1**: (street,streets), (come,comes), (it,its), (project,projects), (investigation,investigations)
- **Set 2**: (few,new), (bright,night), (bedroom,room)

Here, in Set 1 the word pairs differ in the last few phonemes and in Set 2 the word pairs differ in the first few phonemes. To illustrate the relationship between these word pairs, first, the difference vectors
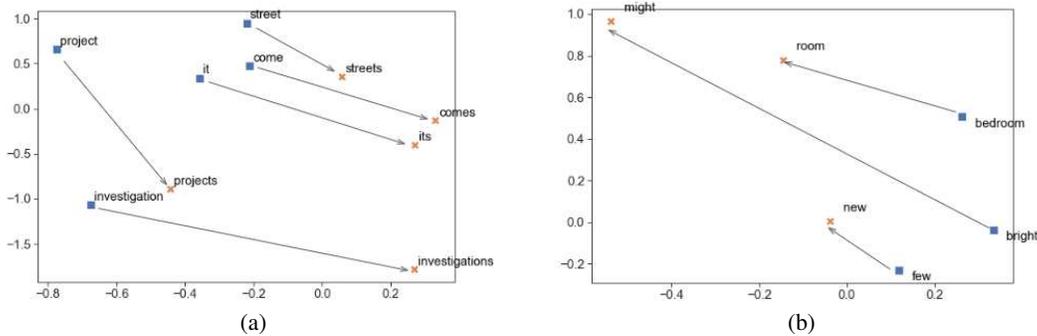
Figure 4: Difference vectors corresponding to (a) Set 1: Word pairs differ in last few phonemes (b) Set 2: Word pairs differ in first few phonemes

were computed between the average spoken-word vector representation of the words present in the above-mentioned word pairs, and then these high dimensional difference vectors were reduced to 2-dimensional vectors using PCA [18], to interpret these vectors. The difference vectors corresponding to Set 1 & Set 2 are shown in Figure 4. It can be observed in the figures that the difference vectors are similar in directions and magnitude. In both the figures, phonetic replacements lead to similar transformations, for example (come → comes) is similar to (it → its) in Figure 4a, and (few → new) is similar to (bright → night) in Figure 4b. These transformations are not perfectly similar because we are taking an average of the same word spoken by different speakers having different accents and pronunciations, but despite this, the transformations are still very close to each other.

All these experiments demonstrate the quality of spoken-word vector representations generated by the proposed model using speech and text entanglement which not only are semantically and syntactically adequate but are also phonetically sound.

6 Conclusion

In this paper, we introduced STEPs-RL for learning phonetically sound spoken-word representations using speech and text entanglement. Our approach achieved an accuracy of 89.47% in predicting phonetic sequences when both gender and dialect of the speaker are used in the auxiliary information. We also compared its performance using different configurations and observed that the performance of the proposed model improved by (1) increasing the spoken word latent representation size, and (2) the addition of auxiliary information like gender and dialect. We were not only able to validate the capability of the learned representations to capture the semantical and syntactical relationships between the spoken-words but were also able to illustrate soundness in the phonetic structure of the generated vector space. For future work, we plan to (1) extend the model to use attention mechanisms, (2) improve performance by using transformer-based architecture, and (3) experimenting on larger datasets, and (4) using features other than MFCCs.

References

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end

speech recognition in english and mandarin. volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

[2] Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 278–289.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[4] Herve A Bourlard and Nelson Morgan. 2012. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media.

[5] Y. Chen, S. Huang, H. Lee, Y. Wang, and C. Shen. 2019. Audio word2vec: Sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9):1481–1493.

[6] Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aaron van den Oord. 2019. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 27(12):2041–2053.

[7] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M. J. F. Gales, K. M. Knill, A. Ragni, H. Wang, and P. Woodland. 2015. Multilingual representations for low resource speech recognition and keyword search. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 259–266.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[9] John S Garofolo, L F Lamel, W M Fisher, Jonathan G Fiscus, D S Pallett, and Nancy L Dahlgren. 1993. Darpa timit acoustic-phonetic continuous speech corpus cd-rom {TIMIT}.

[10] James Glass. 1999. Challenges for spoken dialogue systems. In *Proceedings of the 1999 IEEE ASRU Workshop*, volume 696.

[11] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

[12] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

[13] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414.

[14] Christian Herff and Tanja Schultz. 2016. Automatic speech recognition from neural signals: a focused review. *Frontiers in neuroscience*, 10:429.

[15] Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[17] Takaaki Hori, Jaejin Cho, and Shinji Watanabe. 2018. End-to-end speech recognition with word-based rnn language models. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 389–396. IEEE.

[18] Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

[19] Wei-Ning Hsu, Yu Zhang, and James Glass. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*.

[20] Frederick Jelinek. 1997. *Statistical methods for speech recognition*. MIT press.

[21] H. Kamper. 2019. Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6535–3539.

[22] Sameer Khurana, Antoine Laurent, Wei-Ning Hsu, Jan Chorowski, Adrian Lancucki, Ricard Marxer, and James Glass. 2020. A convolutional deep markov model for unsupervised speech representation learning.

[23] Jinyu Li, Abdelrahman Mohamed, Geoffrey Zweig, and Yifan Gong. 2015. Lstm time and frequency recurrence for automatic speech recognition. In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, pages 187–191. IEEE.

[24] Xiangang Li and X. Wu. 2015. Modeling speaker variability using long short-term memory networks for speech recognition. In *INTERSPEECH*.

[25] Shaoshi Ling, Julian Salazar, Yuzong Liu, and Katrin Kirchhoff. 2020. BERTphone: Phonetically-aware Encoder Representations for Utterance-level Speaker and Language Recognition. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 9–16.

[26] Andy T. Liu, Shang-Wen Li, and Hung yi Lee. 2020. Tera: Self-supervised learning of transformer encoder representation for speech.

[27] Yajie Miao, Mohammad Gowayyed, and Florian Metze. 2015. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE.

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[29] Yasufumi Moriya and Gareth JF Jones. 2018. Lstm language model adaptation with images and titles for multimedia automatic speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 219–226. IEEE.

[30] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

[31] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[32] Linda Polka, Adriel John Orena, Megha Sundara, and Jennifer Worrall. 2017. Segmenting words from fluent speech during infancy–challenges and opportunities in a bilingual context. *Developmental science*, 20(1):e12419.

[33] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath. 2019. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

[34] Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE.

[35] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. 2018. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102.

[36] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

[37] M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

[38] Tian Tan, Yanmin Qian, Dong Yu, Souvik Kundu, Liang Lu, K. Sim, Xiong Xiao, and Yu Zhang. 2016. Speaker-aware training of lstm-rnns for acoustic modelling. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5280–5284.

[39] Zhiyuan Tang, Ying Shi, Dong Wang, Yang Feng, and Shiyue Zhang. 2017. Memory visualization for gated recurrent neural networks in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2736–2740. IEEE.

[40] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. 2013. The second 'chime'speech separation and recognition challenge: An overview of challenge systems and outcomes. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 162–167. IEEE.

[41] Dongqiang Yang and David Martin Powers. 2006. *Verb similarity on the taxonomy of WordNet*. Masaryk University.

[42] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. 2017. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2466. IEEE.