

AUDIO-LINGUISTIC EMBEDDINGS FOR SPOKEN SENTENCES

Albert Haque¹ Michelle Guo¹ Prateek Verma² Li Fei-Fei¹

¹Department of Computer Science, Stanford University

²Center for Computer Research in Music and Acoustics, Stanford University

ABSTRACT

We propose spoken sentence embeddings which capture both acoustic and linguistic content. While existing works operate at the character, phoneme, or word level, our method learns long-term dependencies by modeling speech at the sentence level. Formulated as an audio-linguistic multitask learning problem, our encoder-decoder model simultaneously reconstructs acoustic and natural language features from audio. Our results show that spoken sentence embeddings outperform phoneme and word-level baselines on speech recognition and emotion recognition tasks. Ablation studies show that our embeddings can better model high-level acoustic concepts while retaining linguistic content. Overall, our work illustrates the viability of generic, multi-modal sentence embeddings for spoken language understanding.

1. INTRODUCTION

During verbal communication, humans process multiple words sequentially, often waiting for a full sentence to be completed. Yet, many written and spoken language systems depend on individual character and word-level representations either implicitly or explicitly. This lack of sentence-level context can make it difficult to understand sentences containing conjunctions [1], negations [2], and vocal pitch variations [3]. In this work, we investigate embeddings in the domain of spoken language processing and propose spoken sentence embeddings, capable of modeling both acoustic and linguistic content in a single latent code.

Machine representation of words dates back to ASCII. This one-hot representation encodes each character using a mixture of dummy or indicator variables. While this was slowly extended to words, the large vocabulary size of languages made it difficult. Learned, or distributed, word vector representations [4] replaced one-hot encodings. These word vectors are able to capture semantic information including context from neighboring words. Even today, the community continues to build better contextual word embeddings such as ELMo [5], ULMFit [6], and BERT [7]. Word, phoneme, and grapheme embeddings like Speech2Vec [8] and Char2Wav [9] have also been proposed for speech, following techniques from natural language understanding.

While word-level embeddings are promising, they are often insufficient for speech-related tasks for several reasons. First, word and phoneme embeddings capture a narrow temporal context, often a few hundred milliseconds at most. As a result, these embeddings cannot capture long-term dependencies required for higher-level reasoning (e.g., paragraph or song-level understanding). Almost all of the systems for speech recognition focus on the correctness of local context (e.g., letters, words, and phonemes) rather than overall semantics. Second, for speech recognition, an external language model is often used to correct character and word-level predictions. This requires the addition of complex, multiple hypothesis generation methods [10].

Sentence-level embeddings offer advantages over word and character embeddings. A sentence-level embedding can capture latent factors across words. This is directly useful for higher-level audio tasks such as emotion recognition, prosody modeling, and musical style analysis. Furthermore, most external language models operate at the sentence-level. By having a single sentence-level embedding, the embedding can capture both acoustic and linguistic content at longer contextual window sizes – thus alleviating the need for an external language model entirely, by learning the temporal structure.

Contributions. In this work, our contributions are two-fold. First, we propose moving from phoneme, character, and word-level representations to sentence-level understanding by learning spoken sentence embeddings. Second, we design this embedding to capture both linguistic and acoustic content in order to learn latent codes which can be applicable to a variety of speech and language tasks. We verify the quality of the embedding in our ablation studies, where we assess the generality of sentence-level embeddings when used for automatic speech recognition and emotion classification. We believe this work will inspire future work in speech processing, semantic understanding, and multi-modal transfer learning.

2. METHOD

Our method allows us to learn spoken sentence embeddings that capture both acoustic and linguistic content. In this section, we discuss (i) how we handle long sequences with a temporal convolutional network [11] and (ii) how to learn audio-linguistic content under a multitask learning framework [12].

2.1. Temporal Convolutional Network (TCN)

Our goal is to learn a spoken sentence embedding, which can be used for a variety of speech tasks. Recurrent models are often the default starting point for sequence modeling tasks [11]. For most applications, the state-of-the-art approach to start with is very often a recurrent model. This is evident in machine translation [13, 14], automatic speech recognition [15, 16, 17], and speech synthesis tasks [18, 19, 9].

However, recurrent models such as recurrent neural networks (RNNs) are notoriously difficult to train [20]. For years, machine learning researchers have tried to make RNNs easier to train through novel training strategies [21] and architectures [22]. In [11], the authors show that fully convolutional networks can outperform recurrent networks, without the training complexities [11]. In addition, they can better capture long term dependencies [23] required for such tasks. Motivated by these findings, in this work, we opt for a fully convolutional sequence model (Figure 1). Similar to WaveNet [23], we use a temporal convolutional network (TCN) [11]. While we use the TCN in this work, any causal model will suffice (e.g., Transformer [24]).

Causal Convolutions. To begin, we introduce some quick notation. The sequence modeling task is defined as follows. Given an input sequence x of length T , we have $x = x_1, \dots, x_T$, where each x_t is an observation for timestep t . Suppose we wish to make a prediction y_t at each timestep, then we have $y = y_1, \dots, y_T$. The *causal constraint* states that when predicting y_t , it should depend only on past observations $x_{<t}$ and not future observations. For example, a bidirectional RNN does not satisfy this constraint.

Dilated Convolutions. Standard convolutions have a fixed filter size and thus have a fixed temporal understanding. If our goal is to learn sentence embeddings, we need filters capable of modeling longer temporal windows – ideally modeling the entire sentence. Following the work of [23, 11], we employ dilated convolutions that enable an exponentially large temporal context window at different layers of the TCN. For an input sequence $x = x_1, \dots, x_T \in \mathbb{R}^T$ of length T and a filter $f : \{0, \dots, k-1\} \rightarrow \mathbb{R}$ for some filter size k , the dilated convolution operation F on element s of the sequence is:

$$F(s) = (x \otimes_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{<(T-di)} \quad (1)$$

where d is the dilation factor and \otimes denotes the convolution operator. When $d = 1$, a dilated convolution reduces to a standard convolution. Figure 1 shows a TCN with dilation factor $d = 2$ and kernel size $k = 2$. The TCN allows us to model entire sentences, typically with hundreds to thousands of timesteps, with a single encoder. This encoder produces a fixed-size embedding vector which contains both audio and linguistic information from the input. We explain this process more in the next section.

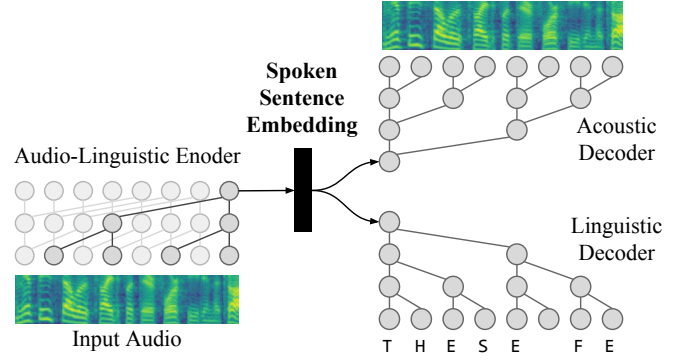


Fig. 1: Audio-linguistic embedding for spoken sentences. Formulated as a multitask learning problem, our method learns a multi-modal spoken sentence embeddings by reconstructing linguistic and acoustic features during training.

2.2. Multitask Learning of Acoustics and Linguistics

In the previous section, we explained how to convert a variable-length spoken sentence into a single, fixed-length vector using a TCN. We now discuss how to encode both acoustic and linguistic content into this vector using multitask learning.

A *task* is a specific prediction problem (e.g., speech recognition, emotion recognition, speaker verification). *Multitask* learning attempts to train a model to simultaneously perform multiple tasks. The simplest method is to combine the task-specific loss functions \mathcal{L}_k into a single loss criterion $\mathcal{L}_{\text{Total}} = \sum_{k=1}^K \lambda_k \mathcal{L}_k$, where λ_k is an arbitrary task-weighting coefficient, where k refers to a task index. We use hard parameter sharing [25] as the TCN encoder is shared for both the acoustic and linguistic decoders. The acoustic target is the original mel-spectrogram input, whereas the linguistic target is a text transcription.

3. EXPERIMENTS

Our experimental procedure consists of two parts. First, we learn individual phoneme, word, and sentence embeddings from TIMIT and LibriSpeech. These include previously published baselines and our proposed spoken sentence embeddings. Second, we evaluate these embeddings on automatic speech recognition and emotion recognition.

3.1. Configuration

Datasets. We train our model on the LibriSpeech dataset [29]. The training set consists of 460 hours of 16 kHz English speech. For speech recognition, we use both LibriSpeech and the TIMIT dataset [30] and report results on the test set. For emotion recognition, we use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [31]. It consists of 24 speakers demonstrating 7 emotions.

Table 1: Comparison of phoneme, word, and sentence embeddings. Sentence embeddings are our proposed method. Phoneme and word error rates are denoted as PER and WER, respectively. All numbers denote percentages. Rows 5 and 6 do not have fusion because the model produces a sentence embedding. Plus/minus denotes 95% confidence interval.

#	Embedding Level	Fusion Method	Speech Recognition		Emotion Recognition		
			TIMIT PER	LibriSpeech WER	Accuracy	Precision	Recall
1	Phoneme	Average [26]	50.4 ± 6.0	74.8 ± 9.6	20.8 ± 5.4	19.5 ± 1.8	19.9 ± 1.9
2	Phoneme	Deep Avg Network [27]	50.8 ± 6.5	70.5 ± 9.3	21.1 ± 5.0	17.5 ± 6.0	21.2 ± 5.0
3	Word [4, 8]	Average [26]	49.6 ± 7.0	30.8 ± 6.0	21.9 ± 3.0	20.3 ± 2.3	21.5 ± 4.0
4	Word [4, 8]	Deep Avg Network [27]	51.2 ± 5.8	18.5 ± 3.0	21.2 ± 4.3	23.0 ± 6.6	23.6 ± 3.7
5	Sentence - RNN (LSTM) [28]		29.6 ± 4.0	10.8 ± 1.8	25.7 ± 5.5	21.0 ± 5.0	24.0 ± 6.0
6	Sentence - Fully Convolutional [11]		30.1 ± 4.5	13.5 ± 2.0	29.2 ± 3.1	28.8 ± 5.4	29.2 ± 3.1

Because no official train-test split is provided, results are reported using four-fold cross-validation (75% train, 25% test). The input representation used in all methods are log-mel spectrograms with 80 mel filters, with audio sampled at 16kHz.

Evaluation Metrics. Speech recognition performance is evaluated using the phoneme (PER) and word error rate (WER). Emotion recognition, a multi-class classification problem, is evaluated using accuracy, precision, and recall.

3.2. Baselines

We now discuss the baselines we compare our method against. To compute a sentence-level embedding, one must decide (i) which intermediate embedding to use and (ii) how to fuse them into a single, sentence-level embedding.

3.2.1. Intermediate Embedding

We select two baseline embeddings for learning intermediate speech representations. The embeddings were selected due to their ease of use and applicability to language tasks.

1. *Speech2Vec* [8]. The Speech2Vec method learns an embedding for spoken words. Word alignments (i.e., segmentations) are required, but the method can be trained in an unsupervised manner. Different utterances of the same word will have different embeddings. This is the same as Word2Vec [4], but instead applied to spoken language.
2. *Phoneme2Vec*. This is the same as Speech2Vec but applied to phonemes. Each utterance of a phoneme is encoded as a single embedding. This is useful for fine-grained tasks.

We train Speech2Vec and Phoneme2Vec on LibriSpeech, similar to the procedure in [8]. The word and phoneme alignments are computed using a Gaussian mixture model [32].

3.2.2. Fusion Method

For these baselines, once we have embeddings for each word or phoneme, we must combine them into a single sentence embedding. We evaluate two fusion methods:

1. *Uniform Average* [26]. The intermediate embeddings are converted to a sentence embedding by computing an element-wise sum of the intermediate embeddings at each word or phoneme position [26] and dividing by the number of words in the sentence. This is a uniform average.
2. *Deep Averaging Network (DAN)* [27]. First, a uniform average is computed over the intermediate embeddings. Then, the averaged vector is fed into a deep neural network to produce the final sentence embedding. This neural network is trained with the encoder, using the same multitask loss objective from Section 2.2.

For DAN, a deep neural network converts the simple average into a sentence embedding. This neural network is the encoder component of a larger encoder-decoder model [13] which is trained offline. This encoder-decoder model is formulated as a multitask learning problem with two decoders: one for acoustic and one for semantic (or linguistic) content.

4. RESULTS

4.1. Speech Processing Tasks

After training our model on LibriSpeech, we take the encoder and use that to generate spoken sentence embeddings on TIMIT [30] and RAVDESS [31]. The weights of encoder are fixed after training and then is only used to extract embeddings. We now have sentence embeddings for TIMIT and RAVDESS and can proceed to train a simple RNN decoder for automatic speech recognition and a SVM classifier for emotion recognition. Table 1 shows results for both tasks.

Sentence-level embeddings outperform phoneme and word embeddings on both tasks. This is true for both the RNN and our fully convolutional TCN model. Phoneme-level embeddings perform poorly for speech recognition (WER in excess of 70%) in contrast to the embedding learned by our method. Comparing the fusion methods, the deep averaging network [27] did not significantly outperform uniform average. The uniformly averaged vector (i.e., input) was already the information bottleneck.

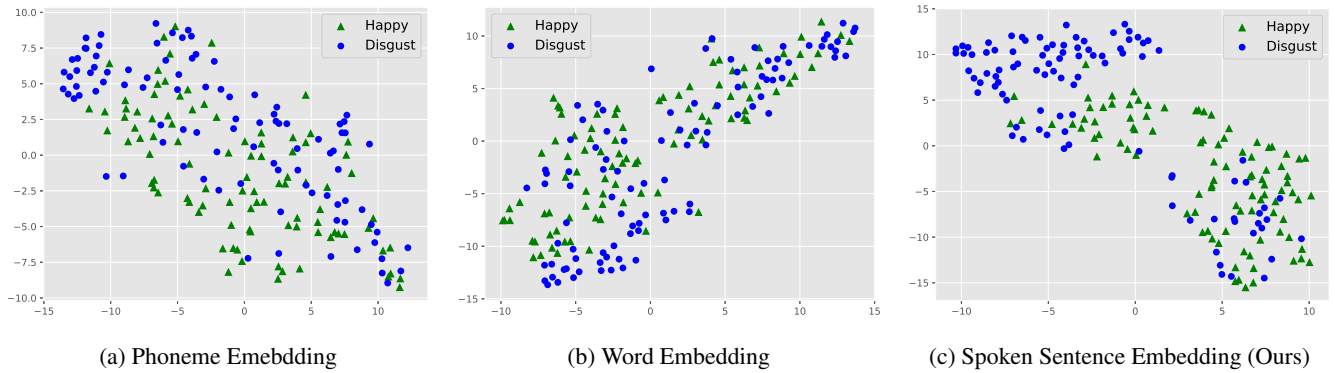


Fig. 2: Visualization of how our method can separate emotions. Each point denotes an audio clip. The x and y axes denote T-SNE projection axes [33]. First, the model was trained on LibriSpeech for speech recognition. Second, the model was used as a fixed feature extractor on the RAVDESS emotion dataset. Finally, the features were plotted and colored by label.

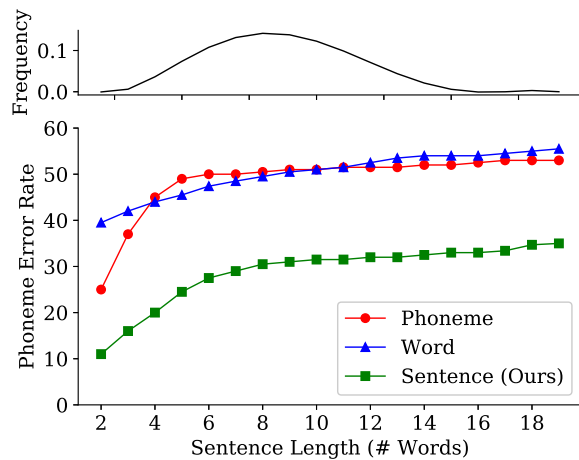


Fig. 3: Effect of sequence length on TIMIT. (Top) Distribution of sentence lengths. (Bottom) Speech recognition phoneme error rate for different embeddings.

4.2. Ablation Studies

Visualization of the Embedding Space. For this experiment, we took our pre-trained spoken sentence encoder, trained on LibriSpeech, and extracted sentence embeddings for each RAVDESS [31] audio clip. We then visualized the embeddings for different emotions in Figure 2. Note that our model never saw a training example from RAVDESS. Without being trained on emotion recognition, our sentence embedding can still cluster emotions.

Sequence Length vs Performance. Figure 3 shows the effect of sequence length on speech recognition performance. In general, as the sentence becomes longer, the performance of the simple embeddings demonstrate lower performance. Our method also demonstrates such a pattern (albeit, to a lesser extent), because longer sentences have more content to “fit” into the fixed-size embedding. This problem could be remedied with variable-length embeddings, which grow linearly in size with the length of the sentence.

5. DISCUSSION

Related Work. The concept of sentence representations (and longer [34]) has been explored in both the natural and spoken language processing communities. On the natural language side, skip-thought vectors [35] and paraphrastic sentence embeddings [36] have been proposed. On the speech side, whole sentence maximum entropy models [37], whole sentence models [38], and time delay networks [39] have been proposed. Closely related to our work is Speech2Vec [8], which was used in our experiments. While Speech2Vec is designed for words, it could be extended to sentences as well.

Most similar to our work is the Universal Sentence Encoder (USE) from natural language processing [26]. They propose two sentence encoders. Each encoder accepts a sentence and produces a single vector. The first encoder is based on attention [24]. The second encoder averages individual word embeddings and feeds the result into a neural network [27]. In contrast to USE, our method does not model intermediate word-level features but instead directly learns a sentence embedding. We believe this can better model long-term context as inter-word relationships are not lost through averaging.

Conclusion & Future Work. In this work, we presented a method for learning spoken sentence embeddings which capture both acoustic and linguistic content. We formulated the problem as a multi-task learning problem to reconstruct both acoustic and linguistic information. Our results show that our spoken sentence embeddings can be used for emotion recognition and speech recognition. Future work can focus on learning “universal” embeddings for spoken language. There has been promising work on the natural language side [26], of which speech and audio can provide additional dimensions. Similar to document-level embeddings for text [34], further audio research could explore larger temporal context sizes for full-length songs and multi-sentence recordings. Overall, our work illustrates the viability of generic, multi-modal sentence embeddings for spoken language understanding.

6. REFERENCES

- [1] Nikita Kitaev and Dan Klein, “Constituency parsing with a self-attentive encoder,” *ACL*, 2018.
- [2] Hao Li and Wei Lu, “Learning with structured representations for negation scope extraction,” in *ACL*, 2018.
- [3] Sylvie Mozziconacci, “Pitch variations and emotions in speech,” in *International Congress of Phonetic Sciences*, 1995.
- [4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS*, 2013.
- [5] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, “Deep contextualized word representations,” *NAACL*, 2018.
- [6] Jeremy Howard and Sebastian Ruder, “Universal language model fine-tuning for text classification,” in *ACL*, 2018.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv*, 2018.
- [8] Yu-An Chung and James Glass, “Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech,” *Interspeech*, 2018.
- [9] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, “Char2wav: End-to-end speech synthesis,” 2017.
- [10] Dario Amodei et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *ICML*, 2016.
- [11] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *arXiv*, 2018.
- [12] Rich Caruana, “Multitask learning,” *Machine Learning*, 1997.
- [13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, “Sequence to sequence learning with neural networks,” in *NIPS*, 2014.
- [14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” *ICLR*, 2015.
- [15] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *ICASSP*, 2016.
- [16] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, “Speech recognition with deep recurrent neural networks,” in *ICASSP*, 2013.
- [17] Alex Graves and Navdeep Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *ICML*, 2014.
- [18] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., “Tacotron: Towards end-to-end speech synthesis,” *Interspeech*, 2017.
- [19] Jonathan Shen et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *ICASSP*, 2018.
- [20] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, “On the difficulty of training recurrent neural networks,” in *ICML*, 2013.
- [21] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in *NIPS*, 2015.
- [22] Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber, “A clockwork rnn,” *arXiv*, 2014.
- [23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv*, 2016.
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *NIPS*, 2017.
- [25] Sebastian Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv*, 2017.
- [26] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al., “Universal sentence encoder,” *arXiv*, 2018.
- [27] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III, “Deep unordered composition rivals syntactic methods for text classification,” in *ACL*, 2015.
- [28] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural Computation*, 1997.
- [29] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*, 2015.
- [30] John S Garofolo, “Timit acoustic phonetic continuous speech corpus,” *Linguistic Data Consortium*, 1993.
- [31] Steven R Livingstone and Frank A Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PloS One*, 2018.
- [32] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, “Montreal forced aligner: trainable text-speech alignment using kaldia,” in *Interspeech*, 2017.
- [33] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *JMLR*, 2008.
- [34] Quoc Le and Tomas Mikolov, “Distributed representations of sentences and documents,” in *ICML*, 2014.
- [35] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, “Skip-thought vectors,” in *NIPS*, 2015.
- [36] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu, “Towards universal paraphrastic sentence embeddings,” *ICLR*, 2016.
- [37] Stanley F Chen and Ronald Rosenfeld, “Efficient sampling and feature selection in whole sentence maximum entropy language models,” in *ICASSP*, 1999.
- [38] Yinghui Huang, Abhinav Sethy, Kartik Audhkhasi, and Bhuvana Ramabhadran, “Whole sentence neural language models,” in *ICASSP*, 2018.
- [39] Vijayaditya Peditinti, Daniel Povey, and Sanjeev Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Interspeech*, 2015.