Improving Unsupervised Acoustic Word Embeddings using Speaker and Gender Information

Lisa van Staden and Herman Kamper
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
18245471@sun.ac.za, kamperh@sun.ac.za

***Abstract*—For many languages, there is little or no labelled speech data available for training speech processing models. In zero-resource settings where unlabelled speech audio is the only available resource, speech applications for search, discovery and indexing often need to compare speech segments of different durations. *Acoustic word embeddings* are fixed dimensional representations of variable length speech sequences, allowing for efficient comparisons. Unsupervised acoustic word embedding models often still retain nuisance factors such as a speaker's identity and gender. Here we investigate how to improve the invariance of unsupervised acoustic embeddings to speaker and gender characteristics. We assume that speaker and gender labels are available for the untranscribed training data. We then consider two different methods for normalising out these factors: speaker and gender conditioning, and adversarial training. We apply both methods to two unsupervised embedding models: a recurrent neural network (RNN) autoencoder and a RNN correspondence autoencoder. In a word discrimination task, we find little benefit by explicitly normalising the embeddings to speaker and gender on English data. But on Xitsonga, substantial improvements are achieved. We speculate that this is due to the higher number of speakers present in the unlabelled Xitsonga training data.**

I. Introduction

Recent research in speech technology has started to consider how systems can be developed in the absence of any transcribed speech resources [1]–[4]. The field of developing speech models solely from unlabelled speech data is referred to as *zero-resource speech processing*. A number of different applications have been developed in this area, including query-by-example search, where the goal is to search over utterances for a given spoken query [5], and unsupervised term discovery (UTD), where the goal is to discover reoccurring speech patterns in a set of untranscribed speech [6]. These type of applications require comparing speech segments of variable length. The conventional method used to do this is dynamic time warping (DTW) which involves finding optimal alignments between speech segments. This method, however, has known limitations, including being computationally expensive [7].

Recent studies have therefore started to explore methods for finding *acoustic word embeddings* of variable length speech segments in a fixed dimensional space [8]–[15]. The idea is that these acoustic embeddings should capture and condense the information in a speech segment that is useful for downstream speech processing tasks. Since arbitrary length speech segments are represented in a fixed-dimensional space, comparing segments becomes a simple distance calculation between embeddings. Many of these studies focus specifically on the zero-resource setting, proposing and investigating different unsupervised methods [10], [12], [13].

The acoustic properties of speech across different speakers and people of different genders[1] vary dramatically. Since the acoustic embeddings, in our case, are learned from unlabelled speech, these properties could still be captured to a large extent in the embeddings. This can lead to a scenario where embeddings for different words from the same speaker can be more similar than embeddings representing the same word from different speakers, and similarly for embeddings from people of different genders. We therefore refer to these properties as nuisance factors. Building on the work of [10], we investigate methods to make unsupervised acoustic word embeddings more invariant to these nuisance factors.

We consider two encoder-decoder recurrent neural network models: the encoder-decoder autoencoder (AE-RNN), first introduced in [13], and the encoder-decoder correspondence autoencoder (CAE-RNN), introduced in [10]. The input to these models are speech segments encoded as sequences of mel-frequency cepstral coefficients (MFCCs). The AE-RNN is trained to reconstruct a input sequence based on a latent variable produced by the encoder. Rather than reconstructing the input itself, the CAE-RNN is trained to reconstruct another instance of the same word. Since the training data is unlabelled, these input-pairs for the CAE needs to be obtained in an unsupervised manner. For this purpose, we use a UTD system which automatically finds similar patterns in the unlabelled audio [16]. For both the AE-RNN and CAE-RNN, the intermediate latent variable is then used as an acoustic word embedding.

To improve invariance to speaker and gender identity, we assume that these properties have been annotated in an unlabelled speech training set. Coarse annotations such as speaker identity and gender would presumably be much easier to obtain than full transcriptions. Using speaker and gender labels, we consider two approaches to improve robustness in unsupervised acoustic word embedding models: conditioning the decoder component of the AE-RNN or CAE-RNN on a

[1] In this paper we make use of the term *gender* instead of *sex* in order to be consistent with the terminology used in the released corpora.

978-1-7281-4162-6/20/$31.00 © 2020 IEEE

trained speaker/gender embedding, and using an adversarial approach where an additional loss encourages the intermediate representation to be a poor signal for speaker or gender classification.

We evaluate the different approaches on two languages by measuring the intrinsic quality of the acoustic word embeddings from each model in a word discrimination task. The speaker and gender information retained in the embeddings are analysed by training separate speaker and gender classifiers on top of the trained embeddings and evaluating the resulting speaker/gender classification accuracy. We show that in both the AE-RNN and CAE-RNN, conditioning models on speaker and gender information or using adversarial training leads to a reduction in some of the information captured by the acoustic word embeddings. However, for the English dataset, the intrinsic quality of the embeddings are only marginally improved. The intrinsic quality of Xitsonga embeddings show greater improvement. We speculate that this is due to the larger number of speakers occurring in the unlabelled training data for this language. Of the two normalisation approaches, we find that adversarial training produces better results for the AE-RNN, but speaker/gender conditioning produced better results for the CAE-RNN. We also find that the approaches are complimentary and can be combined.

II. Proposed Methodology

A. Encoder-Decoder Recurrent Neural Networks

An encoder-decoder recurrent neural network architecture is used as base for our models [17]. The encoder and decoder parts each consist of a stack of recurrent neural networks (RNN). The encoder maps input sequences of variable length into a fixed dimensional latent variable. This latent variable could be the last hidden vector of the last RNN, but in our case we add a linear layer after the encoder to transform the last hidden vector into the latent variable. The decoder then maps the latent variable to an output sequence. In our models, we use stacks of three gated recurrent unit (GRU) networks for both the encoder and decoder [18]. We use one linear layer after the encoder to produce the latent vector, and then another linear layer after the latent vector which then feeds into the decoder. The latent vector is used as the acoustic word embedding.

We experiment with an encoder-decoder autoencoder RNN (AE-RNN) and an encoder-decoder correspondence autoencoder RNN (CAE-RNN). The AE-RNN is trained to map a sequence to a latent vector, which in turn is used to condition the decoder which is trained to reconstruct the original input sequence [13], [19]. Let the AE-RNN model be denoted by f_{AE}. For a given input sequence $X = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T_X}$, where T_X is the length of the input sequence, the reconstructed sequence is $X' = f_{\text{AE}}(X)$, where $X' = \mathbf{x'}_1, \mathbf{x'}_2, ..., \mathbf{x'}_{T_X}$. The loss function applied to one input sequence is as follows:

$$L_{\text{AE}} = ||X - X'||^2 \tag{1}$$

Instead of reconstructing the input itself, the CAE-RNN is trained to reconstruct another instance of the same word as the input. Since our training data is unlabelled, we use an unsupervised term discovery (UTD) system to automatically discover speech segments which are predicted to be of the same type. The UTD system outputs pairs of sequence (X, Y), predicted of the same type, with X as above and $Y = \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{T_Y}$. T_Y is the length of Y.

For a given pair, the CAE-RNN is trained to map a sequence X to a latent variable and then to map the latent variable to Y [10]. Let the CAE-RNN be denoted by f_{CAE}. Then the decoded sequence is $Y' = f_{\text{CAE}}(X)$, where $Y' = \mathbf{y'}_1, \mathbf{y'}_2, ..., \mathbf{y'}_{T_Y}$. The loss function applied to one input pair is as follows:

$$L_{\text{CAE}} = ||Y - Y'||^2 \tag{2}$$

The intuition behind the CAE-RNN is that the model learns to only encode the information into the latent variable that is common between the input-output speech segments (such as

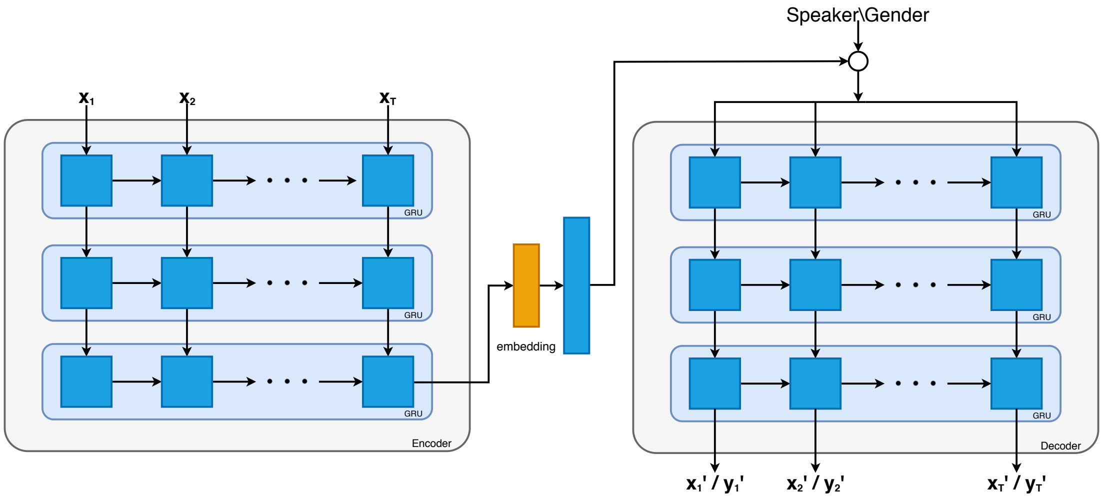

Fig. 1. The architecture of both the AE-RNN and CAE-RNN. The AE-RNN will decode the embedding into X' and the CAE-RNN will decode it into Y'. If the model is conditioned on speaker or gender information, the information is appended to each time step's input for the first GRU in the decoder.

the word identity) while normalising out factors that are not common. The weights of the CAE-RNN are initialised with the weights of a trained AE-RNN. The structure of both the AE-RNN and CAE-RNN is shown in Fig. 1.

B. Reducing Nuisance Factors

There has been a number of studies that propose methods that make the training of speech processing models more robust to nuisance factors [20]–[25]. Speech consists of linguistic and acoustic properties. For a set of utterances consisting of the same word or phrase, but spoken by different speakers, the linguistic content will be the same, but the acoustic properties can differ. We refer to the acoustic properties that relate to factors that are invariant to the linguistic content as nuisance factors. Speaker and gender differences are examples of such nuisance factors.

We investigate two different methods to reduce the speaker and gender information contained in acoustic word embeddings: speaker and/or gender conditioning, and speaker and/or gender adversarial training.

1) Speaker and Gender Conditioning: In the first approach, we hypothesise that conditioning our decoder on the target speaker and/or gender information will make the model less reliant on speaker and/or gender information, specifically in the encoder. This means that the resulting latent variable will (hopefully) be more invariant to the speaker and/or gender information. We create an array of trainable embeddings for each target speaker and gender in our training set. During training, we append this embedding to the decoder input at each time step. Fig. 1 illustrates where in the model the conditioning vector will be added.

2) Adversarial Training: In the second approach, we investigate penalising the model for retaining speaker and gender information inside the acoustic word embedding by adversarially training the models against a speaker or gender classifier. The classifier is a feed-forward neural network (FNN) trained to classify either speaker or gender from an acoustic word embedding. Let N be the number of speakers in the training set or the number of genders and $\mathbf{p} = p_1, ..., p_N$ the set of probabilities for each class returned by the classifier. For a given input latent variable with the true class c, the loss function applied to both classifiers is as follows:

$$L_\text{C} = -p_c + \log(\sum_{i=1}^{N} \log(\exp{(p_i)})) \tag{3}$$

This is an instance of the multiclass log loss. The structure of the classifier can be seen in Fig. 2.

Before the models are adversarially trained against the classifier, we train them using the losses in (1) or (2), respectively, for m epochs, where m is a hyperparameter. Then the classifier is trained on the embeddings produced by the models after m epochs for n epochs, where n is also a hyperparameter. We then set up two turns, turn A and turn B, for each epoch. During turn A, the weights of the classifier are frozen and we train the model (the AE-RNN or CAE-RNN) as usual but also penalise it for speaker or gender information contained in the embedding. During turn B, the weights of theAE-RNN or CAE-RNN model are frozen and we update the weights of the classifier by training it for one epoch on the most recently produced embeddings.

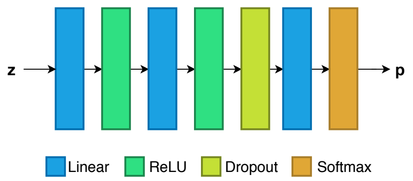

Fig. 2. A FNN that will map an acoustic word embedding, $\mathbf{z}$ to a list of speaker or gender class probabilities, $\mathbf{p}$

The loss function applied to the adversarially trained AE-RNN is as follows:

$$L_\text{AE-Adv} = L_\text{AE} - \gamma L_\text{C} \tag{4}$$

where L_AE is as in (1), L_C is the classification loss and γ is a hyperparameter representing the weight factor of the classification loss.

The loss function applied to the adversarially trained CAE-RNN is as follows:

$$L_\text{CAE-Adv} = L_\text{CAE} - \gamma L_\text{C} \tag{5}$$

where L_CAE is as in (2) and L_C and γ is again a weighing factor. The adversarial process is illustrated in Fig. 3.

III. Experiments

A. Experimental Setup

We train our models on data sets from two languages: English, from the Buckeye corpus [26], and Xitsonga, from the NCHLT corpus [27]. To discover pairs of speech segments we use the UTD-system [16]. This allows our training to remain independent of speech transcription labels. The English training set contains around 14k unique pairs from 12 different

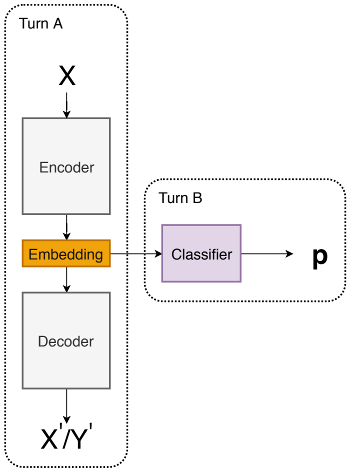

Fig. 3. The adversarial training process consists of two turns, turn A and turn B. During turn A the AE-RNN or CAE-RNN is trained and during turn B, the classifier is trained.

speakers of which 6 are male and 6 are female. The Xitsonga training set contains around 6k unique pairs with 24 different speakers of which 12 are male and 12 are female. The speech segments from the data sets are transformed to sequences of 13-dimensional Mel-frequency cepstral coefficients (MFCCs) with a maximum sequence length of 100.

We follow the model setup of [10]. The dimension of the latent variables is set to 130. All the hidden units in the GRUs of the AE-RNN and CAE-RNN (Fig. 1) have a dimension of 400. We found 50 to be the smallest dimension that achieves the best results for the speaker and gender embeddings. We use learning rates of 0.001 and 0.0001 for the AE-RNN and CAE-RNN, respectively, and both models use a batch size of 256. We use the Adam optimiser [28]. For the English dataset, we train the AE-RNN for 150 epochs and the CAE-RNN for 25 epochs and use early-stopping on validation data. For the Xitsonga dataset, we do not have validation data, so we average the number of epochs that it takes to produce the best models on the English validation data, which is 115 epochs for the AE-RNN and 19 epochs for the CAE-RNN.

During adversarial training, the AE-RNN or CAE-RNN is initially trained for 50 epochs and then the classifier is trained for 100 epochs. For the classifier (Fig. 2) we use a learning rate of 0.001, a batch size of 50 and the Adam optimiser. All three linear layers in the classifier have a dimension 200 and we use a dropout rate of 0.5. During turn A, $\gamma = 0.0001 \times$ the epoch number for the AE-RNN and $\gamma = 0.01 \times$ the epoch number for the CAE-RNN.

B. Evaluation

We want to investigate the speaker and gender information contained in the acoustic word embeddings produced by the AE-RNN and the CAE-RNN and ultimately compare all the embeddings to see if this information is reduced with speaker and/or gender information or with speaker or gender adversarial training. We measure the speaker predictability (SP) and gender predictability (GP) of the embeddings by training a speaker or gender classifier to predict the speaker or gender class of the embedding and evaluating the accuracy. We use the same classifier model as in section II-B2. Note that these models are trained *after* all model training is complete; these classifiers are therefore used as a way to analyse the resulting embeddings to measure the predictability of speaker and gender.

We use the same-different task to evaluate the intrinsic quality of the acoustic word embedding (latent variables) [29]. Two embeddings are similar when the distance between them are less than a certain threshold. By varying thresholds, we create a curve of precision versus the recall. The average precision (AP) of the embeddings is the area under this curve. We consider the AP values to measure the quality of the acoustic word embedding, where higher values are better.

C. Results

We select the two AE-RNN and the two CAE-RNN models that obtained the highest AP scores on the English validation dataset. We evaluate these models on the English and Xitsonga test set along with the baseline AE-RNN and CAE-RNN (the models without any nuisance factor reducing methods

TABLE I
ENGLISH EVALUATION RESULTS

Model		**Conditioning**		**Adversarial**		**Results**		
AE	***CAE***	***Speaker***	***Gender***	***Speaker***	***Gender***	***AP***	***SP***	***GP***
*	-	-	-	-	-	25.19 ± 0.44	74.89 ± 0.21	95.07 ± 0.27
*	-	*	*	-	*	$\mathbf{25.53 \pm 0.46}$	64.36 ± 1.5	89.52 ± 0.27
*	-	-	-	-	*	25.38 ± 0.43	74.89 ± 1.11	93.38 ± 0.47
-	*	-	-	-	-	30.18 ± 0.34	75.59 ± 1.06	93.59 ± 0.6
-	*	*	-	*	-	$\mathbf{30.49 \pm 1.41}$	64.12 ± 0.76	90.14 ± 0.7
-	*	-	*	-	*	29.72 ± 0.76	66.05 ± 1.19	90.51 ± 0.0

TABLE II
XITSONGA EVALUATION RESULTS

Model		**Conditioning**		**Adversarial**		**Results**		
AE	***CAE***	***Speaker***	***Gender***	***Speaker***	***Gender***	***AP***	***SP***	***GP***
*	-	-	-	-	-	11.65 ± 0.34	62.33 ± 0.94	94.23 ± 0.51
*	-	*	*	-	*	$\mathbf{12.78 \pm 1.18}$	48.39 ± 0.39	87.39 ± 0.62
*	-	-	-	-	*	11.22 ± 0.7	60.58 ± 2.82	92.67 ± 0.2
-	*	-	-	-	-	22.52 ± 0.29	54.58 ± 1.0	94.52 ± 0.13
-	*	*	-	*	-	$\mathbf{28.98 \pm 0.36}$	40.28 ± 0.68	88.93 ± 0.53
-	*	-	*	-	*	22.72 ± 1.93	52.57 ± 2.01	89.3 ± 0.51

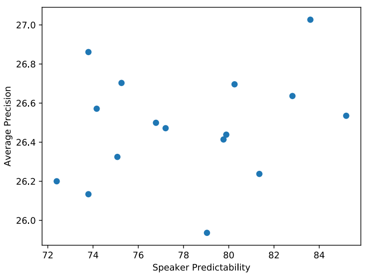

Fig. 4. A scatter plot of the average precision vs. the speaker predictability on the AE-RNN

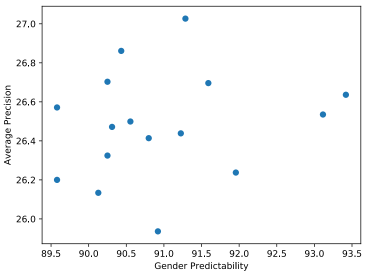

Fig. 5. A scatter plot of the average precision vs. the gender predictability on the AE-RNN

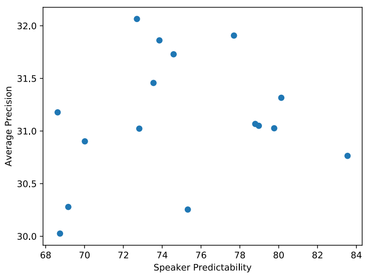

Fig. 6. A scatter plot of the average precision vs. the speaker predictability on the CAE-RNN

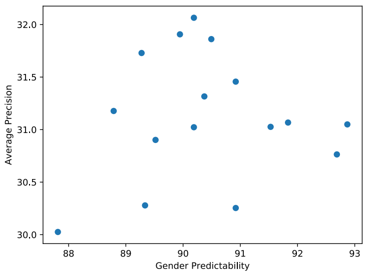

Fig. 7. A scatter plot of the average precision vs. the gender predictability on the CAE-RNN

applied). The results can be seen in Table I and Table II, respectively.

As can be seen in the second and third row of Table I, the nuisance factor normalisation only shows marginally improved scores compared to the baseline AE-RNN, in the first row. Compare the fifth and sixth row with the fourth row, we see that the normalisation also only shows marginal improvement with a combination of speaker conditioning and gender adversarial training and gender conditioning with gender adversarial training shows marginally lower scores compared to the baseline CAE-RNN. The highest AP score was produced by the CAE-RNN with speaker conditioning and speaker adversarial training applied together.

For the Xitsonga data, seen in Table II, when comparing the second row to the baseline AE-RNN in the first row, we see still only small improved score, but the improvement is larger than with the English dataset. The CAE-RNN with speaker conditioning and speaker adversarial training applied together, in the fifth row, scores 22% higehr than the baseline CAE-RNN, in the fourth row. The reason for this model outperforming the baseline model significantly more than with the English model could be because there are twice as many speaker in the Xitsonga training set as in the English training set.

D. *Average Precision and Nuisance Factor Correlation*

We investigate the correlation between the speaker predictability and average precision and the gender predictability and average precision. We analyse the results for all the configurations of both the AE-RNN and CAE-RNN on the English validation dataset. In Fig. 4 and Fig. 5 we see these results on the AE-RNN. It seems like there is no correlation between the AP and SP or GP for the AE-RNN. In Fig. 6 and Fig. 7 we see the result for the CAE-RNN model. It seems like there is potentially a very weak correlation between the AP and SP or GP for the CAE-RNN model. The SP and GP scores for the AE-RNN and CAE-RNN are similar, yet the CAE-RNN shows a slight correlation between AP and SP or GP where the AE-RNN does not.

Taken together, these trends seem to indicate that well-performing embeddings in terms of AP still capture speaker/gender information (at least to a degree where a non-linear classifier can extract this information). Future work will consider a finer-grained analysis of the embeddings.

IV. Conclusion

We have trained encode-decoder autoencoder recurrent neural networks (AE-RNN) and encoder-decoder correspondence autoencoder recurrent neural networks (CAE-RNN) with and without speaker/gender conditioning and with or without a penalising loss term in an adversarial approach. For the English data, applying these nuisance factor reducing methods only shows marginally higher average precision scores. However, the CAE-RNN model trained on Xitsonga with speaker conditioning and speaker adversarial training shows an approximately 22% higher average precision score.

Future work will consider training models on datasets with a larger number of speakers. It will also look at more ways to analyse the acoustic word embeddings, for example, using a linear classifier instead of a non-linear one to measure the speaker and gender predictability of the embeddings.

Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Numbers: 120409).

References

[1] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The Zero Resource Speech Challenge 2015: Proposed Approaches and Results," *Procedia Computer Science*, vol. 81, pp. 67–72, 2016.

[2] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. Black, L. Besacier, S. Sakti, and E. Dupoux, "The Zero Resource Speech Challenge 2019: TTS without T," in *Proc. Interspeech*, 2019.

[3] E. Dunbar, X.-N. Cao Kam, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The Zero Resource Speech Challenge 2017," in *Proc. ASRU*, 2017.

[4] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, M. Seltzer, P. Clark, I. McGraw, B. Varadarajan, E. Bennett, B. Borschinger, J. Chiu, E. Dunbar, A. Fourtassi, D. Harwath, C. Lee, K. Levin, A. Norouzian, V. Peddinti, R. Richardson, T. Schatz, and S. Thomas, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, 2013.

[5] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," in *Proc. Interspeech*, 2017.

[6] A. S. Park and J. R. Glass, "Unsupervised Pattern Discovery in Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[7] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 575–582, 1978.

[8] S. Bengio and G. Heigold, "Word Embeddings for Speech Recognition," in *Proc. Interspeech*, 2014.

[9] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016.

[10] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *Proc. ICASSP*, 2019.

[11] A. L. Maas, S. D. Miller, T. M. ONeil, A. Y. Ng, and P. Nguyen, "Word-level Acoustic Modeling with Convolutional Vector Regression," in *Proc. ICML Workshop Representation Learn*, 2012.

[12] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. ASRU*, 2013.

[13] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio Word2vec: Unsupervised Learning of Audio Segment Representations Using Sequence-to-Sequence Autoencoder," in *Proc. Interspeech*, 2016.

[14] Y.-A. Chung and J. Glass, "Speech2vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech," in *Proc. Interspeech*, 2018.

[15] N. Holzenberger, M. Du, J. Karadayi, R. Riad, and E. Dupoux, "Learning Word Embeddings: Unsupervised Methods for Fixed-size Representations of Variable-length Speech Segments," in *Proc. Interspeech*, 2018.

[16] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," in *arXiv preprint arXiv:1412.3555*, 2014.

[19] K. Cho, B. van Merrinboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation," in *Proc. EMNLP*, 2014.

[20] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, 2013.

[21] Y. Miao, H. Zhang, and F. Metze, "Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 11, pp. 1938–1949, 2015.

[22] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong, and B.-H. Juang, "Speaker-Invariant Training Via Adversarial Learning," in *ICASSP*, 2018.

[23] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1713–1725, 2014.

[24] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant Representations for Noisy Speech Recognition," in *arXiv preprint arXiv:1612.01928*, 2016.

[25] J.-C. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations," in *Proc. Interspeech*, 2018.

[26] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[27] N. Vries, M. Davel, J. Badenhorst, W. Basson, E. Barnard, and A. de Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[28] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2014.

[29] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid Evaluation of Speech Representations for Spoken Term Discovery," in *Proc. Interspeech*, 2011.