MULTILINGUAL ACOUSTIC WORD EMBEDDING MODELS FOR PROCESSING ZERO-RESOURCE LANGUAGES

Herman Kamper[1] *Yevgen Matusevych*[2] *Sharon Goldwater*[2]

[1]E&E Engineering, Stellenbosch University & [2]School of Informatics, University of Edinburgh

kamperh@sun.ac.za, yevgen.matusevych@ed.ac.uk, sgwater@inf.ed.ac.uk

ABSTRACT

Acoustic word embeddings are fixed-dimensional representations of variable-length speech segments. In settings where unlabelled speech is the only available resource, such embeddings can be used in "zero-resource" speech search, indexing and discovery systems. Here we propose to train a single supervised embedding model on labelled data from multiple well-resourced languages and then apply it to unseen zero-resource languages. For this transfer learning approach, we consider two multilingual recurrent neural network models: a discriminative classifier trained on the joint vocabularies of all training languages, and a correspondence autoencoder trained to reconstruct word pairs. We test these using a word discrimination task on six target zero-resource languages. When trained on seven well-resourced languages, both models perform similarly and outperform unsupervised models trained on the zero-resource languages. With just a single training language, the second model works better, but performance depends more on the particular training–testing language pair.

***Index Terms*—** Acoustic word embeddings, multilingual models, zero-resource speech processing, query-by-example.

1. INTRODUCTION

Current automatic speech recognition (ASR) systems use supervised models trained on large amounts of transcribed speech audio. For many low-resource languages, however, it is difficult or impossible to collect such annotated resources. Motivated by the observation that infants acquire language without hard supervision, studies into "zero-resource" speech technology have started to develop unsupervised systems that can learn directly from unlabelled speech audio [1–3].

For zero-resource tasks such as query-by-example speech search, where the goal is to identify utterances containing a spoken query [4, 5], or full-coverage segmentation and clustering, where the aim is to tokenise an unlabelled speech set into word-like units [6–8], speech segments of different durations need to be compared. Alignment methods such as dynamic time warping are computationally expensive and have limitations [9], so *acoustic word embeddings* have been proposed as an alternative: a variable-duration speech segment is mapped to a fixed-dimensional vector so that instances of the same word type have similar embeddings [10]. Segments can then easily be compared by simply calculating a distance between their vectors in this embedding space.

Several supervised and unsupervised acoustic embedding methods have been proposed. Supervised methods include convolutional [11–13] and recurrent neural network (RNN) models [14–17], trained with discriminative classification and contrastive losses. Unsupervised methods include using distances to a fixed reference set [10] and unsupervised autoencoding RNNs [18–20]. The recent unsupervised RNN of [21], which we refer to as the correspondence autoencoder RNN (CAE-RNN), is trained on pairs of word-like segments found in an unsupervised way. Unfortunately, while unsupervised methods are useful in that they can be used in zero-resource settings, there is still a large performance gap compared to supervised methods [21]. Here we investigate whether supervised modelling can still be used to obtain accurate embeddings on a language for which no labelled data is available.

Specifically, we propose to exploit labelled resources from languages where these are available, allowing us to take advantage of supervised methods, but then apply the resulting model to zero-resource languages for which no labelled data is available. We consider two multilingual acoustic word embedding models: a discriminative classifier RNN, trained on the joint vocabularies of several well-resourced languages, and a multilingual CAE-RNN, trained on true (instead of discovered) word pairs from the training languages. We use seven languages from the GlobalPhone corpus [22] for training, and evaluate the resulting models on six different languages which are treated as zero-resource. We show that supervised multilingual models consistently outperform unsupervised monolingual models trained on each of the target zero-resource languages. When fewer training languages are used, the multilingual CAE-RNN generally performs better than the classifier, but performance is also affected by the combination of training and test languages.

This study is inspired by recent work showing the benefit of using multilingual bottleneck features as frame-level representations for zero-resource languages [23–26]. In [27], multilingual data were used in a similar way for discovering acoustic units. As in those studies, our findings show the

Code: https://github.com/kamperh/globalphone_awe

advantage of learning from labelled data in well-resourced languages when processing an unseen low-resource language—here at the word rather than subword level. Our work also takes inspiration from studies in multilingual ASR, where a single recogniser is trained to transcribe speech from any of several languages [28–31]. In our case, however, the language on which the model is applied is never seen during training; our approach is therefore a form of transfer learning [32].

2. ACOUSTIC WORD EMBEDDING MODELS

For obtaining acoustic word embeddings on a zero-resource language, we compare unsupervised models trained on within-language unlabelled data to supervised models trained on pooled labelled data from multiple well-resourced languages. We use RNNs with gated recurrent units [33] throughout.

2.1. Unsupervised monolingual acoustic embeddings

We consider two unsupervised monolingual embedding models. Both have an encoder-decoder RNN structure: an encoder RNN reads an input sequence and sequentially updates its internal hidden state, while a decoder RNN produces the output sequence conditioned on the final encoder output [34, 35].

The unsupervised autoencoding RNN (AE-RNN) of [18] is trained on unlabelled speech segments to reproduce its input, as illustrated in Figure 1. The final fixed-dimensional output $\boldsymbol{z}$ from the encoder (red in Figure 1) gives the acoustic word embedding. Formally, an input segment $X = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ consists of a sequence of frame-level acoustic feature vectors $\boldsymbol{x}_t \in \mathbb{R}^D$ (e.g. MFCCs). The loss for a single training example is $\ell(X) = \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{f}_t(X)||^2$, with $\boldsymbol{f}_t(X)$ the t^{th} decoder output conditioned on the latent embedding $\boldsymbol{z}$. As in [19], we use a transformation of the final hidden state of the encoder RNN to produce the embedding $\boldsymbol{z} \in \mathbb{R}^M$.

We next consider the unsupervised correspondence autoencoder RNN (CAE-RNN) of [21]. Since we do not have access to transcriptions, an unsupervised term discovery (UTD) system is applied to an unlabelled speech collection in the target zero-resource language, discovering pairs of word segments predicted to be of the same unknown type. These are then presented as input-output pairs to the CAE-RNN, as shown in Figure 1. Since UTD is itself unsupervised, the overall approach is unsupervised. The idea is that the model's embeddings should be invariant to properties not common to the two segments (e.g. speaker, channel), while capturing aspects that are (e.g. word identity). Formally, a single training item consists of a pair of segments (X, X'). Each segment consists of a unique sequence of frame-level vectors: $X = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ and $X' = \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{T'}$. The loss for a single training pair is $\ell(X, X') = \sum_{t=1}^{T'} ||\boldsymbol{x}'_t - \boldsymbol{f}_t(X)||^2$, where X is the input and X' the target output sequence. We first pretrain the model using the AE loss above before switching to the CAE loss.

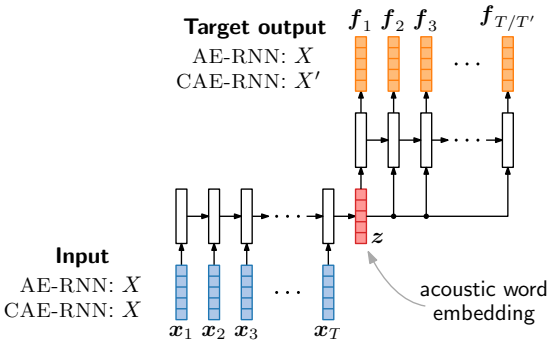

Fig. 1. The AE-RNN is trained to reconstruct its input X (a speech segment) from the latent acoustic word embedding $\boldsymbol{z}$. The CAE-RNN uses an unsupervised term discovery system to find word pairs (X, X'), and is then trained to reconstruct one segment when presented with the other as input.

2.2. Supervised multilingual acoustic embeddings

Given labelled data from several well-resourced languages, we consider supervised multilingual acoustic embedding models.

Instead of using discovered word segments in the target zero-resource language, multilingual AE-RNN and CAE-RNN models can be trained on the pooled ground truth word segments from forced alignments in the well-resourced languages. Although these models are not explicitly discriminative, they do make use of ideal information and are therefore supervised.

As an alternative, we consider an explicitly discriminative model. Given a true word segment X from any one of the training languages, the CLASSIFIERRNN predicts the word type of that segment. Formally, it is trained using the multiclass log loss, $\ell(X) = -\sum_{k=1}^{K} y_k \log f_k(X)$, where K is the size of the joint vocabulary over all the training languages, y_k is an indicator for whether X is an instance of word type k, and $\boldsymbol{f}(X) \in [0, 1]^K$ is the predicted distribution over the joint vocabulary. An acoustic word embedding $\boldsymbol{z}$ is obtained from an intermediate layer shared between all training languages. This embedding is fed into a softmax layer to produce $\boldsymbol{f}(X)$, as illustrated in Figure 2. Embeddings can therefore be obtained for speech segments from a language not seen during training.

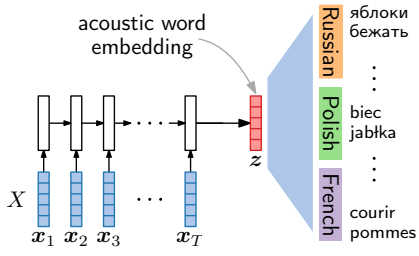

Fig. 2. The multilingual CLASSIFIERRNN is trained jointly on all the training languages to classify which word type an input segment X belongs to. Our model is trained on data from seven languages (three shown here for illustrative purposes).

We could also have used a contrastive loss, but the classifier performs only slightly worse in the supervised monolingual case [11, 14]. It is also easier to extend as it does not require a procedure for sampling pairs over multiple languages.

3. EXPERIMENTAL SETUP

We perform our experiments on the GlobalPhone corpus of read speech [22]. As in [25], we treat six languages as our target zero-resource languages: Spanish (ES), Hausa (HA), Croatian (HR), Swedish (SV), Turkish (TR) and Mandarin (ZH). Each language has on average 16 hours of training, 2 hours of development and 2 hours of test data. Since we do not use transcriptions for the unsupervised monolingual embedding models (§2.1), we apply the UTD system of [36] to each of the training sets, and use the discovered word segments to train unsupervised monolingual CAE-RNN models on each language.[1] Roughly 36k predicted word pairs are extracted in each language. AE-RNN models are similarly trained on the UTD-discovered segments (this gives slightly better performance than training on random segments [21]).

For training supervised multilingual embedding models (§2.2), seven other GlobalPhone languages are chosen as well-resourced languages: Bulgarian (BG), Czech (CS), French (FR), Polish (PL), Portuguese (PT), Russian (RU) and Thai (TH). Each language has on average 21 hours of labelled training data. We pool the data from all languages and train supervised multilingual variants of the AE-RNN and CAE-RNN using true word segments and true word pairs obtained from forced alignments. Rather than considering all possible word pairs from all languages when training the multilingual CAE-RNN, we sample 300k true word pairs from the combined data. Using more pairs did not improve development performance, but increased training time. The multilingual CLASSIFIERRNN is trained jointly on true word segments from the seven training languages. The number of word types per language is limited to 10k, giving a total of 70k output classes (more classes did not give improvements).

Since development data would not be available in a zero-resource language, we performed development experiments on labelled data from yet another language: German. We used this data to tune the number of pairs for the CAE-RNN, the vocabulary size for the CLASSIFIERRNN and the number of training epochs. Other hyperparameters are set as in [21].

All models are implemented in TensorFlow. Speech audio is parametrised as $D = 13$ dimensional static Mel-frequency cepstral coefficients (MFCCs). We use an embedding dimensionality of $M = 130$ throughout, since downstream systems such as the segmentation and clustering system of [8] are constrained to embedding sizes of this order. All encoder-decoder models have 3 encoder and 3 decoder unidirectional RNN layers, each with 400 units. The same encoder structure is used for the CLASSIFIERRNN. Pairs are presented to the CAE-RNN models in both input-output directions. Models are trained using Adam optimisation with a learning rate of 0.001.

We want to measure intrinsic acoustic word embedding quality without being tied to a particular downstream system architecture. We therefore use a word discrimination task designed for this purpose [37]. In the *same-different* task, we are given a pair of acoustic segments, each a true word, and we must decide whether the segments are examples of the same or different words. To do this using an embedding method, a set of words in the test data are embedded using a particular approach. For every word pair in this set, the cosine distance between their embeddings is calculated. Two words can then be classified as being of the same or different type based on some threshold, and a precision-recall curve is obtained by varying the threshold. The area under this curve is used as final evaluation metric, referred to as the average precision (AP). We use the same specific evaluation setup as in [25].

As an additional unsupervised baseline embedding method, we use downsampling [20] by keeping 10 equally-spaced MFCC vectors from a segment with appropriate interpolation, giving a 130-dimensional embedding. Finally, we report same-different performance when using dynamic time warping (DTW) alignment cost as a score for word discrimination.

4. EXPERIMENTAL RESULTS

Table 1 shows the performance of the unsupervised monolingual and supervised multilingual models applied to test data from the six zero-resource languages. Comparing the unsupervised techniques, the CAE-RNN outperforms downsampling and the AE-RNN on all six languages, as also shown in [21] on English and Xitsonga data. It even outperforms DTW on Spanish, Croatian and Swedish; this is noteworthy since DTW uses full alignment to discriminate between words (i.e. it has access to the full sequences without any compression).

Comparing the best unsupervised model (CAE-RNN) to the supervised multilingual models (last two rows), we see that the multilingual models consistently perform better across all six zero-resource languages. The relative performance of the supervised multilingual CAE-RNN and CLASSIFIERRNN models is not consistent over the six zero-resource evaluation languages, with one model working better on some languages while another works better on others. However, both consistently outperform the unsupervised monolingual models trained directly on the target languages, showing the benefit of incorporating data from languages where labels are available.

We are also interested in determining the effect of using fewer training languages. Figure 3 shows development performance on a single evaluation language, Croatian, as supervised models are trained on one, three and all seven well-resourced languages. We found similar patterns with all six zero-resource languages, but only show the Croatian results here. In general the CAE-RNN outperforms the CLASSIFIERRNN when fewer

[1] For extracting pairs, we specifically use the code and steps described at: `https://github.com/eginhard/cae-utd-utils`.

Table 1. AP (%) on test data for the zero-resource languages. The unsupervised CAE-RNNs are trained separately for each zero-resource language on segments from a UTD system applied to unlabelled monolingual data. The multilingual models are trained on ground truth word segments obtained by pooling labelled training data from seven well-resourced languages.

Model	ES	HA	HR	SV	TR	ZH
Unsupervised models:						
DTW	29.7	20.1	13.7	24.2	11.9	27.1
DOWNSAMPLE	19.4	10.7	11.2	16.6	8.0	20.2
AE-RNN (UTD)	18.1	6.5	10.4	12.0	6.8	18.5
CAE-RNN (UTD)	39.7	17.8	21.4	25.2	10.7	21.3
Multilingual models:						
CAE-RNN	**56.0**	**32.7**	29.9	**36.7**	20.9	34.2
CLASSIFIERRNN	54.3	29.5	**32.9**	33.5	**21.2**	**34.5**

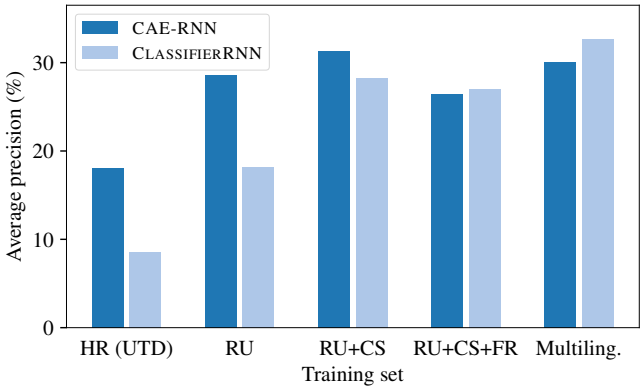

Fig. 3. AP on Croatian (HR) development data for CAE-RNN and CLASSIFIERRNN models as more training languages are added. The 'multiling.' models are trained on all seven well-resourced languages. Scores when training on UTD segments (extracted from unlabelled HR data) are given as a reference.

training languages are used. Using labelled data from a single training language (Russian) already outperforms unsupervised monolingual models trained on Croatian UTD segments. Adding an additional language (Czech) further improves AP. However, when adding French (RU+CS+FR), performance drops; this was not the case on most of the other zero-resource languages, but is specific to Croatian. Croatian is in the same language family as Russian and Czech, which may explain this effect. When all languages are used (right-most bar), the multilingual CAE-RNN performs similarly to the RU+CS case, while the CLASSIFIERRNN performs slightly better.

To further investigate the impact of training language choice, we train supervised monolingual CAE-RNN models on each of the training languages, and then apply each model to each of the zero-resource languages. Results are shown in Figure 4. We observe that the choice of well-resourced language can greatly impact performance. On Spanish, using

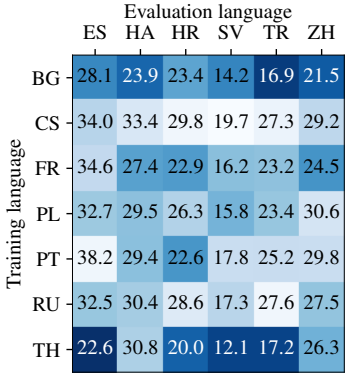

Fig. 4. AP (%) on development data for the six zero-resource languages (columns) when applying different monolingual CAE-RNN models, each trained on labelled data from a well-resourced language (rows). Heatmap colours are normalised for each zero-resource language (i.e. per column).

Portuguese is better than any other language, and similarly on Croatian, the monolingual Russian and Czech systems perform well, showing that training on languages from the same family is beneficial. Although performance can differ dramatically based on the source-target language pair, all of the systems in Figure 4 outperform the unsupervised monolingual CAE-RNNs trained on UTD pairs from the target language (Table 1), except for the BG-HA pair. Thus, training on just a single well-resourced language is beneficial in almost all cases (a very recent study [38] made a similar finding). Furthermore, the multilingual CAE-RNN trained on all seven languages (Table 1) outperforms the supervised monolingual CAE-RNNs for all languages, apart from some systems applied to Turkish. The performance effects of language choice therefore diminish as more training languages are used.

5. CONCLUSION

We proposed to train supervised multilingual acoustic word embedding models by pooling labelled data from well-resourced languages. We applied models trained on seven languages to six zero-resource languages without any labelled data. Two multilingual models (a discriminative classifier and a correspondence model with a reconstruction-like loss) consistently outperformed monolingual unsupervised model trained directly on the zero-resource language. When fewer training languages are used, we showed that the correspondence recurrent neural network outperforms the classifier and that performance is affected by the combination of training and test languages. These effects diminish as more training languages are used. Future work will analyse the types of language-independent properties captured through this transfer learning approach.

This work is based on research supported in part by the National Research Foundation of South Africa (grant number: 120409), a James S. McDonnell Foundation Scholar Award (220020374), an ESRC-SBE award (ES/R006660/1), and a Google Faculty Award for HK.

6. REFERENCES

[1] A. Jansen et al., "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, 2013.

[2] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The Zero Resource Speech Challenge 2015: Proposed approaches and results," in *Proc. SLTU*, 2016.

[3] E. Dunbar et al., "The Zero Resource Speech Challenge 2017," in *Proc. ASRU*, 2017.

[4] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009.

[5] Y.-H. Wang, H.-y. Lee, and L.-s. Lee, "Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection," in *Proc. ICASSP*, 2018.

[6] C.-y. Lee, T. O'Donnell, and J. R. Glass, "Unsupervised lexicon discovery from acoustic input," *Trans. ACL*, vol. 3, pp. 389–403, 2015.

[7] M. Elsner and C. Shain, "Speech segmentation with a neural encoder model of working memory," in *Proc. EMNLP*, 2017.

[8] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," in *Proc. ASRU*, 2017.

[9] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 6, pp. 575–582, 1978.

[10] 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.

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

[12] Y. Yuan et al., "Learning acoustic word embeddings with temporal context for query-by-example speech search," *Proc. Interspeech*, 2018.

[13] A. Haque, M. Guo, P. Verma, and L. Fei-Fei, "Audio-linguistic embeddings for spoken sentences," in *Proc. ICASSP*, 2019.

[14] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *Proc. SLT*, 2016.

[15] Y.-A. Chung and J. R. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," in *Proc. Interspeech*, 2018.

[16] Y.-C. Chen, S.-F. Huang, C.-H. Shen, H.-y. Lee, and L.-s. Lee, "Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval," in *Proc. SLT*, 2018.

[17] S. Palaskar, V. Raunak, and F. Metze, "Learned in speech recognition: Contextual acoustic word embeddings," in *Proc. ICASSP*, 2019.

[18] Y.-A. Chung, C.-C. Wu, C.-H. Shen, and H.-Y. Lee, "Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks," *in Proc. Interspeech*, 2016.

[19] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end ASR-free keyword search from speech," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1351–1359, 2017.

[20] 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.

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

[22] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A multilingual text & speech database in 20 languages," in *Proc. ICASSP*, 2013.

[23] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. SLT*, 2012.

[24] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection," in *Proc. ICASSP*, 2017.

[25] E. Hermann, H. Kamper, and S. J. Goldwater, "Multilingual and unsupervised subword modeling for zero-resource languages," *arXiv preprint arXiv:1811.04791*, 2018.

[26] R. Menon, H. Kamper, J. Quinn, and T. Niesler, "Almost zero-resource ASR-free keyword spotting using multilingual bottleneck features and correspondence autoencoders," *Proc. Interspeech*, 2019.

[27] L. Ondel, H. K. Vydana, L. Burget, and J. Černocký, "Bayesian subspace hidden markov model for acoustic unit discovery," *arXiv preprint arXiv:1904.03876*, 2019.

[28] S. Tong, P. N. Garner, and H. Bourlard, "Multilingual training and cross-lingual adaptation on CTC-based acoustic model," *arXiv preprint arXiv:1711.10025*, 2017.

[29] S. Toshniwal et al., "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*, 2018.

[30] J. Cho et al., "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *Proc. SLT*, 2018.

[31] O. Adams, M. Wiesner, S. Watanabe, and D. Yarowsky, "Massively multilingual adversarial speech recognition," *arXiv preprint arXiv:1904.02210*, 2019.

[32] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2009.

[33] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[34] A. Sperduti and A. Starita, "Supervised neural networks for the classification of structures," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 714–735, 1997.

[35] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014.

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

[37] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. Interspeech*, 2011.

[38] Z. Yang and J. Hirschberg, "Linguistically-informed training of acoustic word embeddings for low-resource languages," *Proc. Interspeech*, 2019.