Multilingual transfer of acoustic word embeddings improves when training on languages related to the target zero-resource language

Christiaan Jacobs, Herman Kamper

E&E Engineering, Stellenbosch University, South Africa

20111703@sun.ac.za, kamperh@sun.ac.za

Abstract

Acoustic word embedding models map variable duration speech segments to fixed dimensional vectors, enabling efficient speech search and discovery. Previous work explored how embeddings can be obtained in zero-resource settings where no labelled data is available in the target language. The current best approach uses transfer learning: a single supervised multilingual model is trained using labelled data from multiple well-resourced languages and then applied to a target zero-resource language (without fine-tuning). However, it is still unclear how the specific choice of training languages affect downstream performance. Concretely, here we ask whether it is beneficial to use training languages related to the target. Using data from eleven languages spoken in Southern Africa, we experiment with adding data from different language families while controlling for the amount of data per language. In word discrimination and query-by-example search evaluations, we show that training on languages from the same family gives large improvements. Through finer-grained analysis, we show that training on even just a single related language gives the largest gain. We also find that adding data from unrelated languages generally doesn't hurt performance.

Index Terms: acoustic word embeddings, zero-resource speech processing, transfer learning, languages of Southern Africa.

1. Introduction

Developing robust speech systems for zero-resource languages—where no transcribed speech resources are available for model training—remains a challenge. Although full speech recognition is not possible in most zero-resource settings, researchers have proposed methods for applications such as speech search [1–3], word discovery [4–7], and segmentation and clustering [8–10], making use of only unlabelled speech audio. Many of these applications require speech segments of different lengths to be compared. This is conventionally done using alignment (e.g. with dynamic time warping). But this can be slow and inaccurate.

Acoustic word embedding (AWE) models map a variable duration speech segment to a fixed dimensional vector [11]. The goal is to map instances of the same word type to similar vectors. Segments can then be efficiently compared by calculating the distance in the embedding space. Given the advantages AWEs have over alignment methods, several AWE models have been proposed [12–23]. Many of these are for the supervised setting, using labelled data to train a discriminative model.

For the zero-resource setting, a number of unsupervised AWE approaches have also been explored, many relying on autoencoder-based neural models trained on unlabelled data in the target language [17, 24–26]. However, there still exists a large performance gap between these unsupervised models and their supervised counterparts [11, 26]. A recent alternative for obtaining AWEs on a zero-resource language is to use multilingual transfer learning [27–31]. The goal is to have the benefits of supervised learning by training a model on labelled data from multiple well-resourced languages, but to then apply the model to an unseen target zero-resource language without fine-tuning it—a form of *transductive* transfer learning [32]. This multilingual transfer approach has been shown to outperform unsupervised monolingual AWE models [29, 33].

Although there is clear benefit in applying multilingual AWE models to an unseen zero-resource language, it is still unclear how the particular choice of training languages affects subsequent performance. Preliminary experiments [29] show improved scores when training a monolingual model on one language and applying it to another from the same family. But this has not been investigated systematically and there are still several unanswered questions: Does the benefit of training on related languages diminish as we train on more languages (which might or might not come from the same family as the target zero-resource language)? When training exclusively on related languages, does performance suffer when adding an unrelated language? Should we prioritise data set size or language diversity when collecting data for multilingual AWE transfer?

We try to answer these questions using a corpus of under-resourced languages spoken in Southern Africa. These languages can be grouped into different families based on their linguistic links. We specifically want to see whether it is beneficial to use closely related languages when training a multilingual model catered to a specific zero-resource language, similar to [34]. We divide the corpus into training and test languages, with (some of) the test languages coming from families that also occur in training. We conduct several experiments where we add data from different language families, and also control for the amount of data per language. AWEs are evaluated in an isolated word discrimination task and in query-by-example (QbE) speech search on full utterances. To our knowledge, only one other study [30] has done AWE-based QbE using multilingual transfer.

Our main findings are as follows. (i) Training a multilingual model using languages that are closely related to the target improves performance. This is true, not just because of the increase in data, but because of language diversity. (ii) When we systematically add training languages, the largest improvements are gained from adding a single related language. Adding more related languages gives small gains. Adding unrelated languages generally gives small or no gains, but also doesn't hurt. (iii) For a target language, the performance of a multilingual model trained on unrelated languages can be matched by a model trained with much less data from multiple related languages.

2. Languages of South Africa

The majority of languages in Africa are considered under-resourced [35]. This includes the eleven official languages of South Africa. As show in Figure 1, nine of these languages belong to the larger Southern Bantu family: isiZulu (Zul), isiXhosa
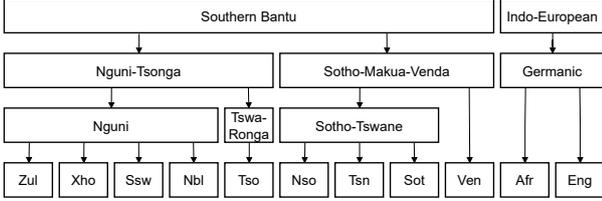
Figure 1: *A family tree for the official South African languages.*[1]

(Xho), Sepedi (Nso), Setswana (Tsn), Sesotho (Sot), Xitsonga (Tso), siSwati (Ssw), Tshivenda (Ven) and isiNdebele (Nbl). Many of these languages are also spoken in countries neighbouring South Africa. From the Southern Bantu family there exist two principal, Nguni-Tsonga and Sotho-Makua-Venda. In the Sotho-Makua-Venda subfamily, Ven is somewhat of a standalone. The other two languages, Afrikaans (Afr) and English (Eng), are Germanic languages from the Indo-European family. All of these languages are considered under-resourced, except for Eng [36]. To give an intuitive idea of how related these languages are, most of the Bantu languages that are grouped together at the lowest level of the hierarchy would (to an extent) be intelligible to a native speaker of another languages in the same group.

3. Acoustic word embedding model

We use the CONTRASTIVERNN AWE model of [33].[2] It performed the best of the model variants considered for multilingual transfer in [33]. The model consists of an encoder recurrent neural network (RNN) that produces fixed-dimensional representations from variable-length speech segments. It is trained to minimise the distance between embeddings from speech segments of the same word type while maximising the distance between embeddings from multiple words of a different type. Let's use $X = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ to denote a sequence of speech features. Formally, given speech segments X_a and X_p containing instances of the same word type and multiple negative examples $X_{n_1}, \dots, X_{n_K}$, the CONTRASTIVERNN produces embeddings $\mathbf{z}_a, \mathbf{z}_p, \mathbf{z}_{n_1}, \dots, \mathbf{z}_{n_K}$ (subscripts indicate anchor, positive and negative, respectively). Each embedding is a fixed dimensional vector $\mathbf{z} \in \mathbb{R}^M$. The model is illustrated in Figure 2. Let $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\|\|\mathbf{v}\|$ denote the cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$. The loss given a positive pair (X_a, X_p) and the set of negative examples is then defined as [37]:

$$J = -\log \frac{\exp\{\text{sim}(\mathbf{z}_a, \mathbf{z}_p)/\tau\}}{\sum_{j \in \{p, n_1, \dots, n_K\}} \exp\{\text{sim}(\mathbf{z}_a, \mathbf{z}_j)/\tau\}}, \quad (1)$$

where τ is a temperature parameter, tuned on development data.

In the zero-resource setting we don't have labelled data in the target language to construct the positive and negative word pairs required for training. We therefore follow the approach of [29], and train a multilingual model on ground truth word pairs (extracted from forced alignments) from a number of languages for which we have labelled data. Subsequently, at test time, we apply the encoder RNN from the multilingual model to extract AWEs for speech from the target zero-resource language.

4. Query-by-example speech search

For evaluating the different AWE models we use an isolated word discrimination task [38]. Recent findings [39] suggest, however,
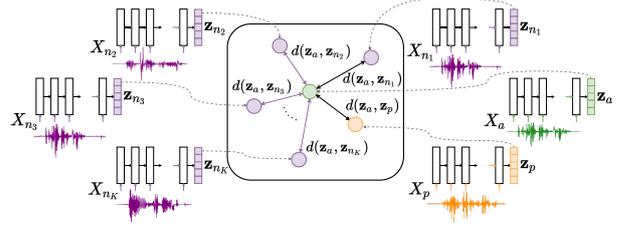

Figure 2: *The CONTRASTIVERNN. The model is trained to minimise the distance between the anchor and positive item $d(\mathbf{z}_a, \mathbf{z}_p)$ while maximising the distance between the anchor and multiple negatives $(\mathbf{z}_a, \mathbf{z}_{n_k})$.*

that this evaluation is not always indicative of downstream system performance. We therefore also perform query-by-example (QbE) speech search, which in contrast to the word discrimination task, does not assume a test set of isolated words, but instead operates on full unsegmented utterances.

Concretely, QbE speech search is the task of identifying the utterances in a speech collection that contain instances of a given spoken query. A number of approaches have been put forward for AWE-based QbE [1, 40, 41]. Here we use the simplified approach from [42]. Using an AWE model, we first embed the query segment. If we knew the word boundaries in the search collection, we could embed each of the words in an utterance and simply look up the closest embeddings to the query. Instead, because we do not have word boundaries, each utterance is split into overlapping segments from some minimum to some maximum duration. Each segment from each utterance is then embedded separately using the AWE model. Finally, to do the QbE task, the query embedding is compared to each of the utterance sub-segment embeddings (using cosine distance), and the minimum distance over the utterance is then taken as the score for whether the utterance contains the given query.

5. Experimental setup

Data. We perform all our experiments on the NCHLT speech corpus [36], which provides wide-band speech from each of the eleven official South African languages (§2). We use a version of the corpus where all repeated utterances were removed, leaving roughly 56 hours of speech from around 200 speakers in each language. We use the default training, validation and test sets. We treat six of the languages as well-resourced: Xho[*], Ssw[*], Nbl[*], Nso[†], Tsn[†] and Eng[‡]. We use labelled data from the well-resourced languages to train a single supervised multilingual model and then apply the model to the target zero-resource languages. More specifically, we extract 100k true positive word pairs using forced alignments for each training language. We select Zul[*], Sot[†], Afr[‡], and Tso as our zero-resource languages and use another language, Ven, for validation of each model.[3] Training and evaluation languages are carefully selected such that for each evaluation language at least one language from the same family is part of the training languages, except for Tso which is in a group of its own.

Models. All speech audio is parametrised as 13-dimensional static Mel-frequency cepstral coefficients. The encoder unit of the CONTRASTIVERNN models (§3) consists of three unidirectional RNNs with 400-dimensional hidden vectors, with an embedding size of $M = 130$ dimensions. Models are optimised using Adam optimisation [43] with a learning rate of 0.001. The

---

[1] https://southafrica-info.com/arts-culture/11-languages-south-africa/

[2] We extend the code available at https://github.com/christiaanjacobs/globalphone_awe_pytorch.

[3] We use superscripts to indicate the different language families: [*]Nguni, [†]Sotho-Tswana, [‡]Germanic.

	Xho*	Zul*	Ssw*	Nbl*	Nso†	Tsn†	Sot†	Afr‡	Eng‡	Tso	Ven
Xho*	84.0	58.5	62.3	55.2	24.2	55.1	43.0	25.8	50.5	55.6	32.3
Zul*	62.9	79.7	62.6	52.8	20.2	47.3	35.8	19.9	34.4	46.8	25.2
Ssw*	45.2	50.0	84.4	51.0	26.0	56.9	44.1	20.1	28.6	53.5	26.0
Nbl*	61.5	58.1	60.7	78.1	22.8	56.9	44.9	23.2	34.8	53.9	32.2
Nso†	43.6	23.2	46.6	31.6	72.4	72.3	61.9	31.8	46.7	60.2	44.9
Tsn†	37.8	27.0	45.5	38.2	43.5	88.8	65.6	30.6	36.9	54.1	38.2
Sot†	33.2	28.0	45.4	38.8	48.5	81.6	83.4	36.4	57.2	56.5	40.7
Afr‡	14.9	10.8	27.4	16.9	12.5	39.5	28.8	90.5	42.7	30.7	14.4
Eng‡	16.9	12.1	26.8	17.5	15.4	45.6	28.9	47.2	95.7	35.7	17.0
Tso	43.0	25.8	44.3	34.2	32.4	57.4	49.8	26.2	31.6	92.1	44.9
Ven	36.3	19.1	33.6	29.9	39.1	62.8	50.5	21.2	29.0	67.5	82.0

Figure 3: *AP (%) when training a monolingual supervised* CON-TRASTIVERNN *on each language (rows) and then evaluating it on each of the other languages (columns). Heatmap colours are normalised for each evaluation language (i.e. per column).*

temperature parameter τ in (1) is set to 0.1.

Evaluation. We consider two tasks for evaluating the performance of the AWE models. First, we use the *same-different* word discrimination task [38] to measure the intrinsic quality of the AWEs. To evaluate a particular AWE model, a set of isolated test word segments is embedded. We use roughly 7k isolated word instances per language from the test data. 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 distance 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 are particularly interested in obtaining embeddings that are speaker invariant. As in [44], we therefore calculate AP by only taking the recall over instances of the same word spoken by different speakers.

The second task is QbE (§4). For each evaluation language we use approximately two hours of test utterances as the search collection. Sub-segments for the utterances in the speech collection are obtained by embedding windows stretching from 20 to 60 frames with a 3-frame overlap. For each evaluation language we randomly draw instances of 15 spoken query word types from a disjoint speech set (the development set—which we never use for any validation experiments) where we only consider query words with at least 5 characters for Afr and Zul and 3 for Sot. There are between 6 and 51 occurrences of each query word. For each QbE test, we ensure that the relevant multilingual AWE model has not seen any of the search or query data during training or validation. We report precision at ten (P@10), which is the fraction of the ten top-scoring retrieved utterances from the search collection that contains the given query.

6. Experimental results

6.1. Cross-lingual evaluation

Before looking at multilingual modelling, we first consider a cross-lingual evaluation where we treat each language as a training language, train a supervised monolingual AWE model, and then apply it to every other language. This allows us to see the effect of training on related languages in a pairwise fashion. The

Table 1: *AP (%) on test data for multilingual models trained on different combinations of well-resourced languages. Models are applied to two zero-resource languages from different language families, Nguni and Sotho-Tswana. For each training language 100k word pairs were extracted.*

Multilingual model	Zul*	Sot†
Nguni:		
Xho* + Ssw* + Nbl*	**68.6**	—
Xho* + Ssw* + Eng‡	60.9	—
Xho* + Nso† + Eng‡	55.7	—
Tsn* + Nso† + Eng‡	37.5	—
Xho* + Ssw* + Nbl* (10% subset)	58.6	—
Sotho-Tswana:		
Nso† + Tsn†	—	**76.7**
Nso† + Eng‡	—	64.8
Xho* + Ssw*	—	51.9
Xho* + Eng‡	—	52.5
Nso† + Tsn† (10% subset)	—	58.4

results are shown in Figure 3. For each evaluation language excluding Tso and Ven, which are in family groups of their own, the best results are achieved from models trained on a language from the same family. E.g. on Zul, Xho is the best training language giving an AP of 58.5%. Eng is the only exception where the model trained on Sot performs better than using Afr, the other Germanic language. Although Ven is in its own group at the lowest layer of the family tree in Figure 1, some of the best results when evaluating on Ven are obtained using models trained on Sotho-Tswana languages (Sot, Tsn, Nso), which are in the same family at a higher level. We also see that for all nine Bantu evaluation languages, worst performance is obtained from the two Germanic models (Afr, Eng).

6.2. Multilingual evaluation

The cross-lingual experiment above was in large part our inspiration for the subsequent analysis of multilingual models. Concretely, we hypothesise that even better performance can be achieved by training on multiple languages from the same family as the target zero-resource language, and that this would be superior to multilingual models trained on unrelated languages. Focusing on two evaluation languages, Zul and Sot from distinct language families, we investigate this hypothesis by training multilingual models with different language combinations, as shown in Table 1. Firstly, we see the best result on a language when all the training languages comes from the same family as the target. Secondly, we see how the performance gradually decrease as the number of training languages related to the target drop. Furthermore, notice the performance boost from including even just one training language related to the target compared to not including any. E.g. on Sot we see a increase of more than 12% absolute when adding just one related language (from 52.5% and 51.9% to 64.8%).

To further demonstrate the benefit of using training languages from the same family, we train a multilingual model for each evaluation language on all its related languages using a 10% subset of the original data. For both Zul and Sot, the subset models outperform the models where no related languages are used. E.g. on Zul, the Xho+Ssw+Nbl subset model outperforms the full Tsn+Nso+Eng model (no related languages) by more than 20% in AP. Moreover, this subset model (58.6% in AP) even outperforms the Xho+Nso+Eng model (55.7%) where
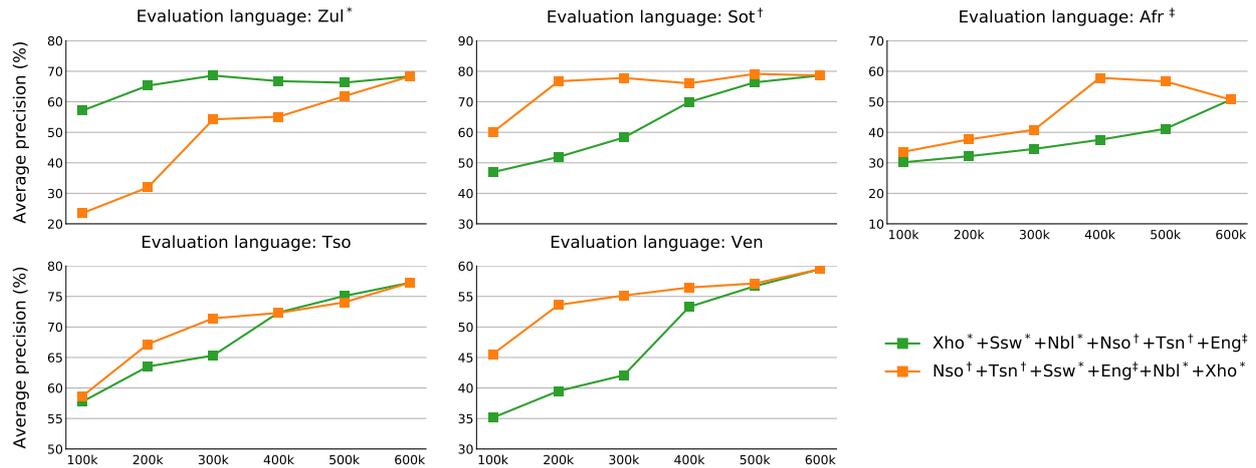
Figure 4: *Same-different results from two sequences of multilingual models, trained by adding one language at a time. For each training language, 100k positive word pairs are used, which is indicated on the x-axis.*
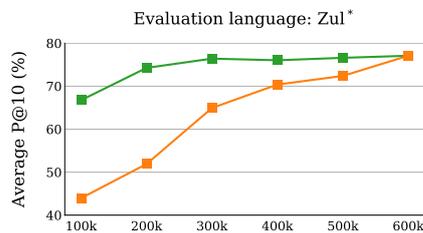

Figure 5: *QbE results on Zul using the same sequences of multilingual models as in Figure 4.*

all the training data from one related language are included and almost matches the AP when using two related languages (Xho+Ssw+Eng, 60.9%).

These comparisons do more than just show the benefit of training on related languages: they also show that it is beneficial to train on a diverse set of related languages.

6.3. Adding more languages

In the above experiments, we controlled for the amount of data per language and saw that training on languages from the same family improves multilingual transfer. But this raises a question: will adding additional unrelated languages harm performance? To answer this, we systematically train two sequences of multilingual models on all six well-resourced languages, evaluating each target language as a new training language is added.

Same-different results for all five evaluation languages are shown in Figure 4 and QbE results for Zul are shown in Figure 5. As for Zul, the trends in the same-different and QbE results track each other closely for the other evaluation languages (so these are not shown here). In the fist sequence of multilingual models (green), we start by adding the three Nguni languages (Xho, Ssw, Nbl), followed by the two Sotho-Tswana languages (Nso, Tsn), and lastly the Germanic training language (Eng). The second sequence (orange) does not follow a systematic procedure.

On Zul, the green sequence, which starts with a related language (Xho), initially achieves a higher score compared to the orange sequence in both Figures 4 and 5. Then, the score gradually increase by adding more related languages (Ssw, Nbl). Thereafter, adding additional unrelated languages (Nso, Tsn, Eng) show no performance increase. In fact, AP decreases slightly after adding the two Sotho-Tswana languages (Nso, Tsn), but not significantly. The orange sequence starts low on

Zul until the first related language (Ssw) is added, causing a sudden increase. Adding the Germanic language (Eng) has little effect. Adding the last two related languages (Nbl, Xho) again causes the score to increase. A similar trend follows for Sot and Afr, where adding related languages causes a noticeable performance increase, especially when adding the first related language; after this, performance seem to plateau when adding more unrelated languages. (Afr is the one exception, with a drop when adding the last language in the orange sequence). On Tso, which does not have any languages from the same family in the training set, AP gradually increases in both sequences without any sudden jumps. Although Ven isn't in the same family as Nso and Tsn at the lowest level of the tree in Figure 1, it belongs to the same family (Sotho-Tswana) at a higher level. This explains why it closely tracks the Sot results.

Summarising these results, we see that adding unrelated languages generally does not decrease scores, but also does not provide a big benefit (except if it is one of the earlier languages in the training sequence, where data is still limited). In contrast, it seems that training on languages from the same family is again beneficial; this is especially the case for the first related language, irrespective of where it is added in the sequence.

7. Conclusion

We investigated the effect of training language choice when applying a multilingual acoustic word embedding model to a zero-resource language. Using word discrimination and query-by-example search tasks on languages spoken in Southern Africa, we showed that training a multilingual model on languages related to the target is beneficial. We observed gains in absolute scores, but also in data efficiency: you can achieve similar performance with much less data when training on multiple languages from the same family as the target. We showed that even including just one related language already gives a large gain. From a practical perspective, these results indicate that one should prioritise collecting data from related languages (even in modest quantities) rather than collecting more extensive datasets from diverse unrelated families, when building multilingual acoustic word embedding models for a zero-resource language.

8. References

[1] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. ICASSP*, 2015.

[2] S.-F. Huang, Y.-C. Chen, H.-y. Lee, and L.-s. Lee, "Improved audio embeddings by adjacency-based clustering with applications in spoken term detection," *arXiv preprint arXiv:1811.02775*, 2018.

[3] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Learning acoustic word embeddings with temporal context for query-by-example speech search," in *Proc. Interspeech*, 2018.

[4] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 186–197, 2008.

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

[6] L. Ondel, H. K. Vydana, L. Burget, and J. Černockỳ, "Bayesian subspace hidden markov model for acoustic unit discovery," in *Proc. Interspeech*, 2019.

[7] O. Räsänen and M. A. C. Blandón, "Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics," *arXiv preprint arXiv:2008.00731*, 2020.

[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] S. Seshadri and O. Räsänen, "Sylnet: An adaptable end-to-end syllable count estimator for speech," *IEEE Signal Proc. Let.*, vol. 26, no. 9, pp. 1359–1363, 2019.

[10] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Proc. Interspeech*, 2020.

[11] 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.

[12] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proc. Interspeech*, 2014.

[13] W. He, W. Wang, and K. Livescu, "Multi-view recurrent neural acoustic word embeddings," in *Proc. ICLR*, 2017.

[14] 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.

[15] 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.

[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] 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.

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

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

[20] Y. Shi, Q. Huang, and T. Hain, "Contextual joint factor acoustic embeddings," in *Proc. SLT*, 2021.

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

[22] S. Settle, K. Audhkhasi, K. Livescu, and M. Picheny, "Acoustically grounded word embeddings for improved acoustics-to-word speech recognition," in *Proc. ICASSP*, 2019.

[23] M. Jung, H. Lim, J. Goo, Y. Jung, and H. Kim, "Additional shared decoder on Siamese multi-view encoders for learning acoustic word embeddings," in *Proc. ASRU*, 2019.

[24] 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.

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

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

[27] M. Ma, H. Wu, X. Wang, L. Yang, J. Wang, and M. Li, "Acoustic word embedding system for code-switching query-by-example spoken term detection," *arXiv preprint arXiv:2005.11777*, 2020.

[28] H. Kamper, Y. Matusevych, and S. J. Goldwater, "Multilingual acoustic word embedding models for processing zero-resource languages," in *Proc. ICASSP*, 2020.

[29] H. Kamper, Y. Matusevych, and S. Goldwater, "Improved acoustic word embeddings for zero-resource languages using multilingual transfer," *IEEE Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1107–1118, 2021.

[30] Y. Hu, S. Settle, and K. Livescu, "Acoustic span embeddings for multilingual query-by-example search," in *Proc. SLT*, 2021.

[31] Y. Hu, S. Settle, and K. Livescu, "Multilingual jointly trained acoustic and written word embeddings," in *Proc. Interspeech*, 2020.

[32] S. Ruder, "Neural transfer learning for natural language processing," Ph.D. dissertation, NUI Galway, Ireland, 2019.

[33] C. Jacobs, Y. Matusevych, and H. Kamper, "Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation," in *Proc. SLT*, 2021.

[34] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 27, pp. 621–630, 2019.

[35] I. Orife, J. Kreutzer, B. Sibanda, D. Whitenack, K. Siminyu, L. Martinus, J. T. Ali, J. Abbott, V. Marivate, S. Kabongo *et al.*, "Masakhane–machine translation for africa," in *Proc. ICLR*, 2020.

[36] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, 2014.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.

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

[39] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating the reliability of acoustic speech embeddings," in *Proc. Interspeech*, 2020.

[40] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *Proc. ICASSP*, 2015.

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

[42] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *Proc. ICASSP*, 2019.

[43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[44] E. Hermann, H. Kamper, and S. J. Goldwater, "Multilingual and unsupervised subword modeling for zero resource languages," *Comput. Speech Language*, vol. 65, 2021.