Interspeech 2022
18-22 September 2022, Incheon, Korea

Integrating Form and Meaning: A Multi-Task Learning Model for Acoustic Word Embeddings

Badr M. Abdullah, Bernd Möbius, Dietrich Klakow

Language Science and Technology (LST), Saarland University, Germany
Saarland Informatics Campus, Germany
{babdullah|moebius|dietrich}@lsv.uni-saarland.de

Abstract

Models of acoustic word embeddings (AWEs) learn to map variable-length spoken word segments onto fixed-dimensionality vector representations such that different acoustic exemplars of the same word are projected nearby in the embedding space. In addition to their speech technology applications, AWE models have been shown to predict human performance on a variety of auditory lexical processing tasks. Current AWE models are based on neural networks and trained in a bottom-up approach that integrates acoustic cues to build up a word representation given an acoustic or symbolic supervision signal. Therefore, these models do not leverage or capture high-level lexical knowledge during the learning process. In this paper, we propose a multi-task learning model that incorporates top-down lexical knowledge into the training procedure of AWEs. Our model learns a mapping between the acoustic input and a lexical representation that encodes high-level information such as word semantics in addition to bottom-up form-based supervision. We experiment with three languages and demonstrate that incorporating lexical knowledge improves the embedding space discriminability and encourages the model to better separate lexical categories.

Index Terms: acoustic word embeddings, form-to-meaning mapping, cognitive modeling, multi-task learning

1. Introduction

The development of robust automatic speech recognition (ASR) systems requires large collections of high-quality transcribed speech, which are only available for a small subset of the world languages. To facilitate access to spoken content for language varieties that are not yet supported by conventional ASR systems, researchers have developed voice-based search applications such as query-by-example (QbE) search [1–3]. These systems rely on vector-space acoustic models that map variable-length spoken word segments onto fixed-size vector representations such that exemplars of the same word are (ideally) projected onto the same vector [4–8]. In the speech technology literature, these fixed-dimensionality vector representations are known as acoustic word embeddings (AWEs). Currently, the top performing and the most efficient models of AWEs are based on deep neural networks (DNNs) [9–12]. Due to the ubiquity of computers that support DNNs coupled with highly-optimized vector-space search algorithms [13], AWEs enable efficient indexing and retrieval of spoken content at an unprecedented scale.

In addition to their applications in speech technology, DNN-based models of AWEs have been adopted as models of human speech processing and analyzed from a cognitively motivated angle in recent studies. For example, it has been shown that AWEs exhibit a human-like word onset bias where distinct words are more likely to be perceived as similar if they begin with the same sound [14]. Furthermore, AWE models have been shown to predict non-native perceptual difficulties in phonetic categorization [15], cross-linguistic effects in auditory lexical processing [16], and the facilitation effect of cross-language similarity on spoken-word processing [17, 18]. In non-native word production, models of AWEs have been reported to capture lexical production patterns of second language (L2) learners [19]. These empirical findings from cognitively motivated, computational word perception and production studies encourage further integration between speech technology and cognitive science.

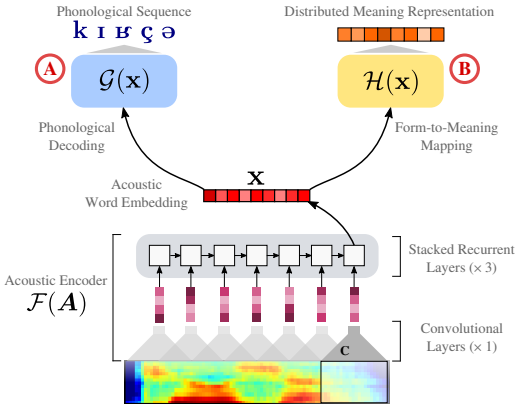

Figure 1: *A schematic view of our proposed model.*

Nevertheless, the majority of existing AWEs rely on supervision signals that only capture low-level, form-based information about the word. That is, AWEs are learned in a bottom-up approach whereby acoustic-phonetic cues are integrated in the model to build up a word form representation that encodes its phonetic features and phonological structure. However, a host of psycholinguistic studies with human listeners have shown that top-down, high-level lexical properties—such as word semantics—not only interact with the word recognition process but also facilitate discrimination between word competitors [20–24]. We take inspiration from these experimental findings and introduce an AWE model based on the multi-task learning framework that integrates form-based and meaning-based supervision signals into a single model (Fig. 1). Contrary to prior work that aims to learn the semantic content directly using a very large speech corpus [25], our model incorporates word semantics as an additional supervision signal, thus requiring only a few hours of speech and being more applicable in low-resource settings. We experiment with read speech corpora for three languages and empirically demonstrate that integrating high-level lexical knowledge into training AWEs improves the ability of the model to discriminate between lexical categories.

10.21437/Interspeech.2022-626

2. AWEs via Multi-Task Learning

Given an acoustic signal that corresponds to a spoken word represented as a temporal sequence of T spectral vectors, i.e., $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T)$, the goal of an AWE model is to transform $\boldsymbol{A}$ into a fixed-dimensionality vector representation $\mathbf{x}$. This task corresponds to learning an encoder function $\mathcal{F}_{\boldsymbol{\theta}} : \mathcal{A} \rightarrow \mathbb{R}^D$, where $\mathcal{A}$ is the (continuous) space of acoustic sequences, D is the dimensionality of the embedding, and $\boldsymbol{\theta}$ are the parameters of the function. Sequences in $\mathcal{A}$ can vary in length, thus the function $\mathcal{F}_{\boldsymbol{\theta}}$ should be modeled with a suitable neural architecture such as recurrent neural networks (RNNs). Therefore, transforming a variable-length acoustic input into a D-dimensional AWE can be described as

$$\mathbf{x} = \mathcal{F}(\boldsymbol{A}; \boldsymbol{\theta}_{\mathcal{F}}) \in \mathbb{R}^D \tag{1}$$

Different approaches in the literature have been proposed for modeling the function $\mathcal{F}(.; \boldsymbol{\theta}_{\mathcal{F}})$, which can be characterized as either architectural innovations or introducing new loss functions. In the approach we propose in this paper, our goal is to integrate two sources of supervision signals—namely phonological form and lexical semantics—into the training procedure. To this end, we assume a dataset $\mathcal{D} = \{(\boldsymbol{A}^1, w^1), (\boldsymbol{A}^2, w^2), \ldots, (\boldsymbol{A}^N, w^N)\}$ of N spoken words where w^i is the written form of the ith word. Such a dataset can be automatically obtained using a forced alignment tool on a transcribed speech dataset. Furthermore, we assume the availability of two look-up dictionaries: (1) a dictionary that maps each written word onto its phonetic transcription as $\Phi(w) = \boldsymbol{\varphi}_{1:\tau} = \{\varphi_1, \varphi_2, \ldots, \varphi_\tau\}$, which can be automatically created using a grapheme-to-phoneme (G2P) tool, and (2) a lookup dictionary that maps each word into a distributed word representation as $\Lambda(w) = \mathbf{w} \in \mathbb{R}^K$. The distributed word representation ideally encodes high-level lexical knowledge about the word—such as its semantic and syntactic properties—and can be obtained independently using a large text corpus or from a public repository of semantic word embeddings such as ***word2vec*** [26], ***Glove*** [27], or ***fasttext*** [28].

2.1. Form-based Phonological Supervision

Our first learning objective is based on the sequence-to-sequence learning framework in which the network is trained as a word-level acoustic model (Fig.1, branch [A]). Given the output of acoustic encoder $\mathbf{x}$, a phonological decoder $\mathcal{G}(.; \boldsymbol{\theta}_{\mathcal{G}})$ aims to decode the corresponding phonological sequence $\boldsymbol{\varphi}_{1:\tau}$ of the word form $\mathbf{x}$. The objective is to minimize a categorical cross-entropy loss at each timestep in the decoder, which is equivalent to minimizing the term

$$\begin{aligned} \mathcal{L}^{\phi}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{G}}) &= -\sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} \log \mathbf{P}\big(\Phi(w^i) \mid \mathbf{x}^i;\, \boldsymbol{\theta}_{\mathcal{G}}\big) \\ &= -\sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} \sum_{t=1}^{\tau} \log \mathbf{P}\big(\varphi_t \mid t,\, \mathbf{x}^i;\, \boldsymbol{\theta}_{\mathcal{G}}\big) \end{aligned} \tag{2}$$

where $\mathbf{P}\big(\varphi_t \mid t,\, \mathbf{x}^i;\, \boldsymbol{\theta}_{\mathcal{G}}\big)$ is the probability of the phoneme φ_t at the tth timestep, conditioned on the previous phoneme sequence $\boldsymbol{\varphi}_{1:t-1}$ and the AWE $\mathbf{x}$, and $\boldsymbol{\theta}_{\mathcal{G}}$ are the parameters of the decoder. The intuition of this learning objective is the following: although their acoustic realizations vary due to speaker and context variability, different exemplars of the same word category would have identical phonetic transcriptions. Therefore, we expect the model to project exemplars of the same lexical category nearby in the embedding space and the distance in embedding space should ideally correlate with phonological (dis)similarity.

2.2. Meaning-based Lexical Supervision

Our second learning objective aims to map the acoustic input $\boldsymbol{A}$ onto a high-level lexical representation (Fig.1, branch [B]). The goal here is to incorporate a supervision signal from a level that is higher in the linguistic hierarchy compared to form-based phonological supervision. Inspired by Maas et al. [29], we model this task as a vector regression problem. The output of the acoustic encoder $\mathbf{x}$ is transformed via a feed-forward network into a semantic vector as $\boldsymbol{v} = \mathcal{H}(\mathbf{x}; \boldsymbol{\theta}_{\mathcal{H}}) \in \mathbb{R}^K$. Thus, the objective is to minimize the term

$$\mathcal{L}^{\lambda}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{H}}) = \sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} || \boldsymbol{v}^i - \Lambda(w^i) ||^2 \tag{3}$$

where $\Lambda(w^i) \in \mathbb{R}^K$ is the ground-truth distributed representation, or semantic word embedding, of the ith sample. We assume that continuous-space, distributed word representations are available to the model during training. Given the ubiquity of semantic word embeddings in the natural language processing research and the availability of large scale text corpora for many languages, we believe that our assumption is reasonable.

2.3. Integrating Form and Meaning Supervision

To integrate the two sources of supervision when training the model, we jointly minimize the term

$$\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}, \boldsymbol{\theta}_{\mathcal{G}}, \boldsymbol{\theta}_{\mathcal{H}}) = \alpha \cdot \mathcal{L}^{\phi} + \beta \cdot \mathcal{L}^{\lambda} \tag{4}$$

Here, α and β are trade-off hyperparameters (i.e., scalars) that control the contribution of each term to the overall loss.

3. Baseline: Contrastive Acoustic Model

We compare the performance of our proposed model to a strong baseline that explicitly minimizes the distance between exemplars of the same lexical category. The baseline model employs a contrastive triplet loss that has been extensively explored in the AWEs literature with different underlying architectures and has shown strong discriminative performance [9, 30–32]. Given a matching pair of AWEs $(\mathbf{x}^a, \mathbf{x}^+)$—i.e., embeddings of two exemplars of the same word type—the objective is then to minimize a triplet margin loss

$$\mathcal{L}(\boldsymbol{\theta}_{\mathcal{F}}) = \sum_{(\boldsymbol{A}^i, w^i) \in \mathcal{D}} \max\big[0, \mu + d(\mathbf{x}^i, \mathbf{x}^+) - d(\mathbf{x}^i, \mathbf{x}^-)\big] \tag{5}$$

where $\mathbf{x}^-$ is an AWE that corresponds to a word other than w^i, and $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, 1]$ is the cosine distance. This objective aims to map acoustic exemplars of the same word closer in the embedding space while pushing away segments of different word types by a distance defined by the margin hyperparameter μ. To obtain negative samples, we use hard negative sampling [33], that is, we create mismatching pairs from the mini-batch such that $d(\mathbf{x}^i, \mathbf{x}^-)$ is minimized.

4. Experiments

4.1. Experimental Data

The data in our study is drawn from the GlobalPhone multilingual speech database [34] for Portuguese, German, and Polish (see Table 1). We sample 42 speakers from each language for training and obtain spoken word segments using the Montreal Forced Aligner [35]. It is worth pointing out that the speakers in the validation and test splits are held-out and not used while

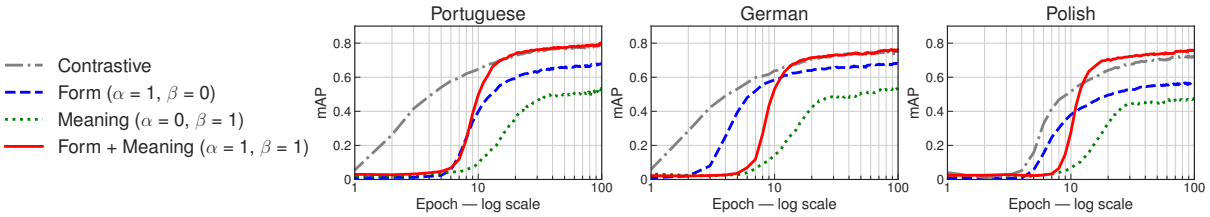

Figure 2: *Learning curves of the models for 100 training epochs, quantified by the word discrimination task and the mAP metric.*

training. The phonetic transcription for each word is produced using the *eSpeak* G2P tool. Then, each acoustic segment is parametrized as a sequence of 39-dimensional Mel-frequency spectral coefficients of 25ms frames with 10ms overlap.

4.2. Architecture and Hyperparameters

Acoustic Encoder $\mathcal{F}(.;\boldsymbol{\theta}_{\mathcal{F}})$. Our acoustic encoder consists of a hybrid, convolutional-recurrent neural network architecture. The frond-end consists of a 1D convolutional layer of 64 filters with a kernel size of 5 spectral vectors and stride of 2. Then, the output of the convolutional layer is fed sequentially into a recurrent block that consists of a 3-layer unidirectional Gated Recurrent Unit (GRU) with a hidden state of 512 units, which yields a 512-dimensional AWE as the last hidden state of the GRU. We apply layer-wise dropout with a probability of 0.2. Bidirectional GRUs did not yield further improvements.

Phonological Decoder $\mathcal{G}(.;\boldsymbol{\theta}_{\mathcal{G}})$. We employ a 1-layer GRU of 512 units hidden state that takes the 512-dimensional AWE as the initial hidden state and decodes the corresponding phonological sequence without teacher forcing.

Form-to-Meaning Regressor $\mathcal{H}(.;\boldsymbol{\theta}_{\mathcal{H}})$. We employ a linear layer ($512 \rightarrow 300$) followed by a `tanh` non-linearity to project the AWE $\mathbf{x}$ onto the corresponding distributed word representation. We use pre-trained 300-dimensional fasttext embeddings as distributed word representations. Deeper feed-forward networks did not yield further improvements.

Contrastive Loss. For the baseline model with the contrastive loss, we experiment with different values of the margin hyperparameter $\mu = \{0.2, 0.3, 0.4, 0.5\}$, out of which 0.4 yields the best performance on the validation set.

Training Details. We train all models in this study for 100 epochs with batches of 256 samples using the Adam optimizer [36] with an initial learning rate (LR) of 0.001. The LR is reduced by a factor of 0.5 if the performance on the validation set does not improve for 10 epochs. The epoch with the best validation performance is used for evaluation on the test set.

Implementation. We develop our code using PyTorch [37] and we make it publicly available (`https://github.com/uds-lsv/semantically_enriched_AWEs`).

Table 1: *Word-level statistics of our experimental data.*

	# segments per split			duration ($\mu \pm \sigma$)	TTR
	train	valid	test		
Portuguese	28810	9029	9580	0.51 ± 0.19	0.147
German	28914	9683	9372	0.44 ± 0.18	0.193
Polish	27979	9656	9089	0.50 ± 0.18	0.267

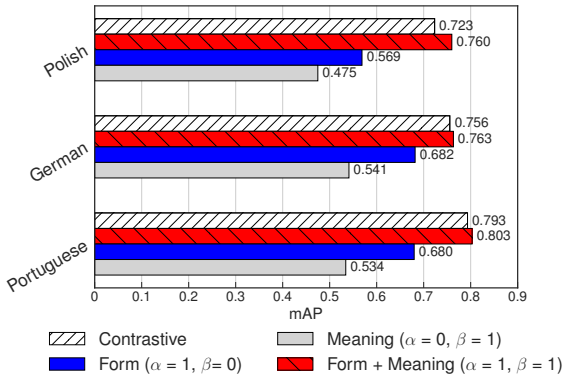

Figure 3: *Word discrimination performance (mAP) on test set.*

4.3. Experimental Results

We conduct an intrinsic evaluation for the AWEs to assess the performance of our models using the same-different acoustic word discrimination task with the mean average precision (mAP) metric [31, 38, 39]. Prior work has shown that performance on this task positively correlates with improvement on downstream QbE speech search [32]. This task evaluates the ability of the model to determine whether two given speech segments correspond to the same word type—that is, whether or not two acoustic segments are exemplars of the same category.

Fig. 2 shows the learning curves for the models during 100 epochs of training quantified by the performance on the validation set. Contrary to the other models, we observe that the contrastive baseline model reaches a reasonable performance before the 10th epoch, which we attribute to the fact that the evaluation task (word discrimination) and the learning objective (contrastive triple loss) are analogous. Fig. 3 shows the final performance on the test set. We observe that both the form-only model ($\alpha = 1, \beta = 0$) and the meaning-only model ($\alpha = 0, \beta = 1$) perform poorly compared to the contrastive baseline. However, integrating the two sources of supervision in the form + meaning setting ($\alpha = 1, \beta = 1$) enables the model to outperform the contrastive baseline for the three languages in our study. The gain in performance is more prominent in the Polish language (relative mAP gain by 5.06%), which is the most morphologically complex language in our study due to its rich inflection system. The Polish morphological complexity is also reflected in its relatively high type-to-token ratio (TTR) in Table 1. These findings show that integrating high-level linguistic knowledge in training acoustic models improves the discriminability of the embeddings space, and the effect seems to be more prominent on a language with a rich morphological system.

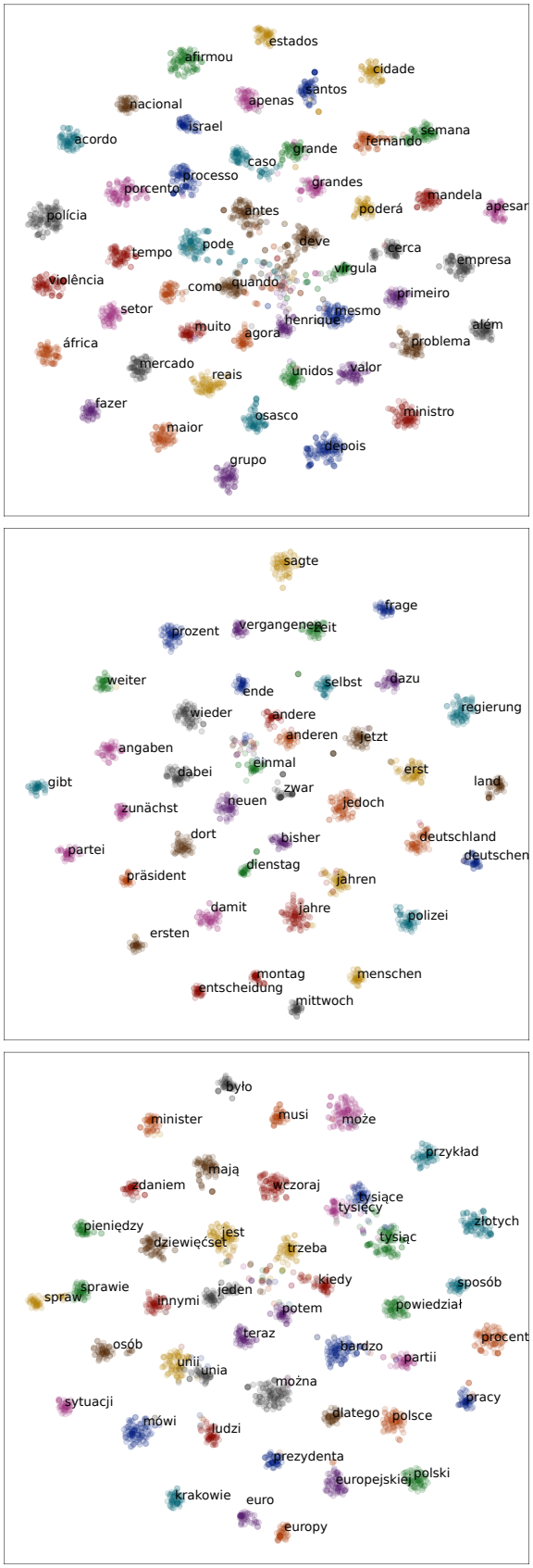

Figure 4: *Two-dimensional visualization of the word embedding space using the t-SNE algorithm: (Top) Portuguese, (Middle) German, and (Bottom) Polish. Figure is best viewed in color.*

4.4. Embedding Visualization

To analyze our multi-task learning model and gain further insights into its emergent embedding space, we use the t-SNE [40] dimensionality reduction algorithm on the AWEs from the setting where $\alpha = 1$ and $\beta = 1$ (form + meaning supervision). We visualize the embeddings of acoustic word samples from held-out speakers (i.e., a set of speakers the models were not trained on). Note that the t-SNE objective aims to preserve the local structure within the higher-dimensional space when reducing the dimensionality. Therefore, the local distance between the two-dimensional projections of the embeddings mainly reflects the cluster structure within the embedding space, which enables us to visually inspect the emergent clusters and investigate whether or not they correspond to distinct lexical categories.

The t-SNE visualizations for the three languages in our study are illustrated in Fig. 4. We observe a clear tendency for exemplars of the same lexical category to closely cluster in the embedding space, despite the lack of an explicit clustering objective in the learning procedure. One notable exception we observe in Fig. 4 (bottom) for the Polish language is the two nearby, nearly overlapping clusters that correspond to the word forms [tysięcy] and [tysiące]. Note that these two word forms are two morphological variants of the same lemma (i.e., [tysiąc], the Polish word for *thousand*). Given their semantic and phonological similarity, a small distance between the centroids of their clusters is expected.

5. Discussion and Conclusion

AWEs are vector representations of spoken words that encode their acoustic-phonetic features and phonological structures. In addition to their utility in speech technology applications, models of AWEs have shown to produce human-like behavior in various auditory lexical processing tasks. Existing methods for learning AWEs from speech corpora employ training strategies with acoustic, phonological feature-based, or symbolic form-based supervision. These learning strategies correspond to the bottom-up integration of acoustic-phonetic cues to build up a word form representation. In our paper, we have introduced a methodology based on the multi-task learning framework that leverages top-down, high-level lexical knowledge to learn semantically-enriched AWEs. We have experimented with semantic word embeddings as distributed meaning representations that guide the learning process in addition to form-based phonological supervision. Our experiments have demonstrated that integrating the two sources of supervision (i.e., phonological form and lexical semantics) improves the discriminability of the embeddings space—for the three languages in our study—as evidenced by the competent performance of our model compared to a strong contrastive AWE model. Furthermore, the t-SNE visualization analysis has supported our experimental findings in the word discrimination evaluation and provided further evidence that incorporating top-down lexical knowledge encourages the model to better separate the lexical categories in the embedding space without explicit supervision that directs the network to minimize the distance in the embedding space or a clustering objective.

6. Acknowledgements

We thank the anonymous reviewers for their feedback and Miriam Schulz for her comments on the paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

7. References

[1] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009.

[2] A. Jansen and B. V. Durme, "Indexing raw acoustic features for scalable zero resource search," in *Proc. Interspeech*, 2012.

[3] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "The spoken web search task at MediaEval 2012," in *Proc. ICASSP*, 2013.

[4] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

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

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

[7] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

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

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

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

[11] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015.

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

[13] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, 2017.

[14] Y. Matusevych, H. Kamper, and S. Goldwater, "Analyzing autoencoder-based acoustic word embeddings," in *BAICS Workshop ICLR*, 2020.

[15] Y. Matusevych, T. Schatz, H. Kamper, N. Feldman, and S. Goldwater, "Evaluating computational models of infant phonetic learning across languages," in *Proc. CogSci*, 2020.

[16] Y. Matusevych, H. Kamper, T. Schatz, N. H. Feldman, and S. Goldwater, "A phonetic model of non-native spoken word processing," in *Proc. EACL*, 2021.

[17] B. Abdullah, I. Zaitova, T. Avgustinova, B. Möbius, and D. Klakow, "How familiar does that sound? Cross-lingual representational similarity analysis of acoustic word embeddings," in *Proc. of BlackboxNLP Workshop, EMNLP*, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.blackboxnlp-1.32

[18] A. Mayn, B. M. Abdullah, and D. Klakow, "Familiar words but strange voices: Modelling the influence of speech variability on word recognition," in *Proc. EACL, Student Research Workshop*, 2021.

[19] S. Ando, N. Minematsu, and D. Saito, "Lexical Density Analysis of Word Productions in Japanese English Using Acoustic Word Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 4433–4437.

[20] J. Zhuang, B. Randall, E. A. Stamatakis, W. D. Marslen-Wilson, and L. K. Tyler, "The interaction of lexical semantics and cohort competition in spoken word recognition: an fMRI study," *Journal of Cognitive Neuroscience*, vol. 23, no. 12, pp. 3778–3790, 2011.

[21] D. Mirman and J. S. Magnuson, "Dynamics of activation of semantically similar concepts during spoken word recognition," *Memory & cognition*, vol. 37, no. 7, pp. 1026–1039, 2009.

[22] E. Strain, K. Patterson, and M. S. Seidenberg, "Semantic effects in single-word naming." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 5, p. 1140, 1995.

[23] M. J. Cortese, G. B. Simpson, and S. Woolsey, "Effects of association and imageability on phonological mapping," *Psychonomic Bulletin & Review*, vol. 4, no. 2, pp. 226–231, 1997.

[24] Y. Hino and S. J. Lupker, "Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 6, p. 1331, 1996.

[25] Y.-A. Chung and J. R. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," in *IProc. Interspeech*, 2020.

[26] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[27] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[28] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1008

[29] A. L. Maas, S. D. Miller, T. M. O'neil, A. Y. Ng, and P. Nguyen, "Word-level acoustic modeling with convolutional vector regression," in *Proc. ICML Workshop Representation Learn*, 2012.

[30] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016.

[31] B. M. Abdullah, M. Mosbach, I. Zaitova, B. Mobius, and D. Klakow, "Do acoustic word embeddings capture phonological similarity? An empirical study," in *Interspeech*, 2021.

[32] C. Jacobs and H. Kamper, "Multilingual transfer of acoustic word embeddings improves when training on languages related to the target zero-resource language," in *Interspeech*, 2021.

[33] A. Jansen, M. Plakal, R. Pandya, D. P. Ellis, S. Hershey, J. Liu, R. C. Moore, and R. A. Saurous, "Unsupervised learning of semantic audio representations," in *Proc. ICASSP*, 2018.

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

[35] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi." in *Interspeech*, 2017.

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

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeuRIPS*, 2019.

[38] R. Algayres, M. S. Zaiem, B. Sagot, and E. Dupoux, "Evaluating the Reliability of Acoustic Speech Embeddings," in *Proc. Interspeech*, 2020.

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

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.