INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study

Badr M. Abdullah[1,2,3], *Marius Mosbach*[1,2,3], *Iuliia Zaitova*[1,2], *Bernd Möbius*[2], *Dietrich Klakow*[1,2,3]

[1]Spoken Language Systems (LSV), Saarland University, Germany
[2]Language Science and Technology (LST), Saarland University, Germany
[3]Saarland Informatics Campus, Germany

{babdullah|mmosbach|izaitova|moebius|dietrich}@lsv.uni-saarland.de

Abstract

Several variants of deep neural networks have been successfully employed for building parametric models that project variable-duration spoken word segments onto fixed-size vector representations, or acoustic word embeddings (AWEs). However, it remains unclear to what degree we can rely on the distance in the emerging AWE space as an estimate of word-form similarity. In this paper, we ask: does the distance in the acoustic embedding space correlate with phonological dissimilarity? To answer this question, we empirically investigate the performance of supervised approaches for AWEs with different neural architectures and learning objectives. We train AWE models in controlled settings for two languages (German and Czech) and evaluate the embeddings on two tasks: word discrimination and phonological similarity. Our experiments show that (1) the distance in the embedding space in the best cases only moderately correlates with phonological distance, and (2) improving the performance on the word discrimination task does not necessarily yield models that better reflect word phonological similarity. Our findings highlight the necessity to rethink the current intrinsic evaluations for AWEs.

Index Terms: acoustic word embeddings, phonological similarity, contrastive learning, deep neural networks

1. Introduction

Spoken language technologies such as spoken term discovery [1, 2, 3] and query-by-example (QbE) search [4, 5, 6] aim to capture, organize, and facilitate access to the linguistic content of spoken documents while abstracting away from speaker- and context-related sources of variability in speech. To this end, researchers have developed parametric models based on deep neural networks (DNNs) that project variable-length spoken word segments onto speaker-invariant vector representations, known as acoustic word embeddings (AWEs), where acoustic segments of the same word are projected nearby in space [7, 8, 9, 10, 11]. AWEs, and their underlying vector-space acoustic models, enable efficient indexing and retrieval of spoken content at a scale that non-parametric template-based approaches with dynamic programming [12, 13] have failed to deliver.

Several DNN architectures and learning objectives have been explored in the literature to build AWEs. State-of-the-art AWE models are trained using either contrastive objectives [14, 15] or reconstruction objectives [16, 17]. AWEs have been used in downstream applications including ASR [18] and QbE search [11, 19]. However, evaluating the utility of AWEs using downstream applications is expensive and may not be always feasible. Therefore, researchers have developed an intrinsic evaluation for AWEs based on the acoustic word discrimination task. In this task, AWE models are evaluated based on their ability to determine whether or not two acoustic segments correspond to the same word type [20, 16, 9].

Furthermore, Levin et al. [7] have hypothesized that the distance in the emergent AWE space can be interpreted as a metric of (perceptual) dissimilarity between linguistic units (e.g., phones, syllables, words). However, none of the previous studies has empirically (in)validated this hypothesis with a rigorous evaluation beyond word discrimination. Although previous studies have proposed to incorporate the pronunciation distance in the learning objective [15, 21], the reported performance showed no improvement on the word discrimination task, while the distance in the AWE space has shown only a weak correlation with orthographic distance [15]. These observations, however, are yet to be systematically investigated across different architectures, objectives, and languages beyond English, which we aim to address in our study.

Since AWE models have been recently adopted as cognitive models of infant phonetic learning [22] and cross-language spoken word processing [23], we argue that more effort should be devoted to analyze and understand the emergent embedding space to make sure it behaves as expected. In this paper, we take a step in this direction and make the following contributions:

(1) We train AWE models with identical resources and hyperparameters and examine the effects of, and the interplay between, the architecture and learning objective on model performance (§2).

(2) We analyze the correlation between the distance in the embedding space and word-form (dis)similarity, which we measure using a phonetically-informed extension of Levenshtein distance (§3 and §4).

(3) We empirically show that while AWE models trained with contrastive objectives outperform other models on the word discrimination task, they are poor at capturing phonological similarity (§5).

2. Acoustic Word Embedding Models

The core component of a neural AWE model is an acoustic encoder, which can be formally described as a parametric 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. Given an acoustic word segment represented as a temporal sequence of T spectral vectors $\overline{\boldsymbol{a}} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T)$, an embedding is computed as $\mathbf{x} = \mathcal{F}_{\boldsymbol{\theta}}(\overline{\boldsymbol{a}}) \in \mathbb{R}^D$. We experiment with convolutional and recurrent architectures for the encoder $\mathcal{F}_{\boldsymbol{\theta}}$ and investigate three learning objectives, which we formally describe below.

http://dx.doi.org/10.21437/Interspeech.2021-678

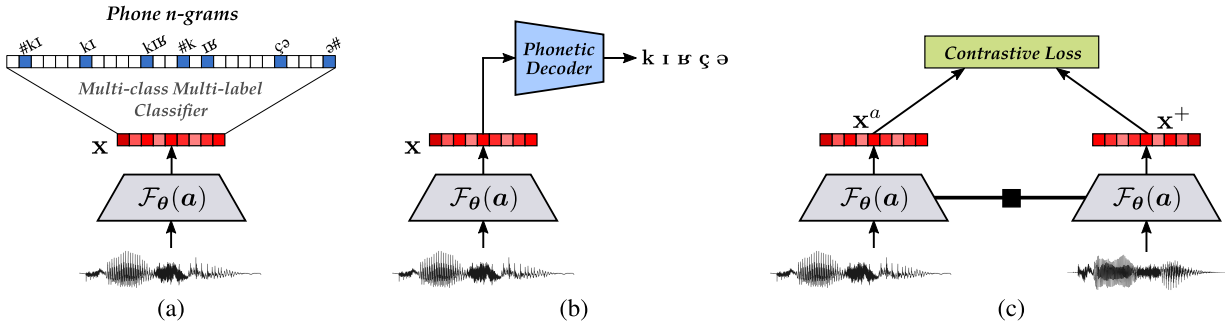

Figure 1: *A visual illustration of the learning objectives for the models in our paper: (a) phone n-gram detection: a classification objective based on a multi-class multi-label classifier, (b) word-to-phones: a sequence-to-sequence objective based on decoding the phonological sequence from the acoustic word embedding, and (c) a contrastive siamese objective with a triplet margin loss.*

2.1. Phone n-gram Detection Objective

Following previous work [24, 25], we use a classification objective as our neural baseline (Fig. 1–a). However, we diverge from previous approaches in the classification target of our network. Instead of predicting the word type, we train our model to detect phone sequences that are present in each acoustic segment. For example, consider acoustic segments that correspond to the word type "*Kirche*". While previous approaches trained the model to predict the word type as an atomic unit, we first transform the phonetic sequence /k ɪ ʁ ç ə/ into a set of phone bigrams {'#k', 'kɪ', 'ɪʁ', 'ʁç', 'çə', 'ə#'} and trigrams {'#kɪ', 'kɪʁ', 'ɪʁç', 'ʁçə', 'çə#'}. We then train the model to predict the presence of each phone n-gram. Formally, given a vocabulary of word types $\mathcal{V}$ where each word is associated with a phonetic sequence $\boldsymbol{\varphi}_{1:N} = (\varphi_1, \ldots, \varphi_N)$, we obtain a set of all phone n-grams in the training data as $\Phi = \bigcup_{v \in \mathcal{V}} \mathcal{N}(\boldsymbol{\varphi}^v_{1:N})$, where $\mathcal{N}$ is a function that converts a phone sequence into phone n-grams. Each phone n-gram in the set Φ is a prediction target in the network where a multi-class multi-label classification head is connected to the AWE to give the prediction output as $\hat{\mathbf{y}} = \sigma(\mathbf{x}\mathbf{W} + \mathbf{b}) \in [0, 1]^{|\Phi|}$, where σ is a sigmoid activation function, $\mathbf{W}$ is a weight matrix, and $\mathbf{b}$ is a bias vector. The objective is to minimize binary cross-entropy loss as

$$\mathcal{L} = -\Big[\mathbf{y} \cdot \log(\hat{\mathbf{y}}) + (\mathbf{1} - \mathbf{y}) \cdot \log(\mathbf{1} - \hat{\mathbf{y}})\Big] \tag{1}$$

where $\mathbf{y} \in \{0, 1\}^{|\Phi|}$ is the ground truth vector and 1 indicates the presence of a phone n-gram in the segment and 0 indicates its absence. This objective has an efficiency advantage over type-based classification approaches since the classification layer is smaller in size, due to the fact that $|\Phi| << |\mathcal{V}|$.

2.2. Word-To-Phones Objective

Our second learning objective is based on sequence-to-sequence learning whereby the network is trained as a word-level acoustic model (Fig. 1–b). Given an acoustic sequence $\overline{\boldsymbol{a}}$ and its corresponding phonetic sequence $\boldsymbol{\varphi}_{1:N}$, the acoustic encoder $\mathcal{F}_{\boldsymbol{\theta}}$ is trained to take as input $\overline{\boldsymbol{a}}$ and produce an AWE $\mathbf{x}$, which is then fed into a recurrent phonetic decoder $\mathcal{G}$ whose goal is to generate the corresponding sequence $\boldsymbol{\varphi}_{1:N}$. The intuition of this objective is that phonologically similar word-forms would usually have overlapping phonetic segments, we thus expect similar words to end up nearby in the embedding space. The objective is to minimize a categorical cross-entropy loss at each timestep in the decoder, which is equivalent to

$$\mathcal{L} = -\sum_{t=1}^{N} \log \mathbf{P}_\theta(\varphi_t | \boldsymbol{\varphi}_{<t}, \mathbf{x}) \tag{2}$$

where $\mathbf{P}_\theta$ is the probability of the phone φ_t at timestep t, conditioned on the previous phone sequence $\boldsymbol{\varphi}_{<t}$ and the AWE $\mathbf{x}$.

2.3. Contrastive Siamese Objective

The third objective we investigate in this paper is the siamese contrastive objective [26], which has been extensively explored in the literature with different underlying architectures [27, 14]. This objective differs from the first two objectives in two aspects: (1) it explicitly minimizes/maximizes relative distances between AWEs of the same/different word types, and (2) models are trained solely on sensory input without symbolic grounding since each word segment is paired with another segment of the same word type. Given a matching pair of AWEs $(\mathbf{x}^a, \mathbf{x}^+)$, the objective is then to minimize a triplet margin loss

$$\mathcal{L} = \max\Big[0, \mu + d(\mathbf{x}^a, \mathbf{x}^+) - d(\mathbf{x}^a, \mathbf{x}^-)\Big] \tag{3}$$

where $\mathbf{x}^-$ is an AWE that corresponds to a different word type, which serves as a negative sample, and $d : \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, 1]$ is the cosine distance. This objective aims to map acoustic segments of the same word type closer in the embedding space while pushing away segments that correspond to other word types by a distance defined by the margin hyperparameter μ. We experiment with two different strategies for choosing the negative sample $\mathbf{x}^-$: (1) we randomly sample a negative AWE from the mini-batch, and (2) we choose the segment that minimizes the distance $d(\mathbf{x}^a, \mathbf{x}^-)$, which is known as semi-hard negative sampling [28]. In §5, we show that the negative sampling strategy has a significant impact on model performance.

3. Phonological Similarity Measure

We adopt the phonologically weighted Levenshtein distance (PWLD) as our measure of phonological distance, or similarity, between different word-forms [29]. The PWLD metric extends the string-based Levenshtein distance (LD) by conditioning the cost of phone substitutions on phonetic similarity, which can be characterized based on the number of distinctive features shared by two phones. PWLD captures, for example, that the German word *sicher* /z ɪ ç ɐ/ is phonologically more similar to *Becher* /b ɛ ç ɐ/ than to *sitzt* /z ɪ ts t/, even though the pairwise LD for

Table 1: *Examples of pairwise word distances with Levenshtein distance (LD) and the phonologically weighted LD (PWLD).*

WORD I		WORD II		LD	PWLD
Orth.	IPA	Orth.	IPA		
sicher	/z ɪ ç ɐ/	*Becher*	/b ɛ ç ɐ/	2	0.263
		Fischer	/f ɪ ʃ ɐ/	2	0.368
		Lichter	/l ɪ ç t ɐ/	2	0.632
		sitzt	/z ɪ ts t/	2	0.795

both pairs is 2 (see Table 1 for more examples). However, we make three adaptations to the original PWDL to make it suitable for our study: (1) we represent every phone in our inventory as a discrete, multi-valued feature vector based on the PHOIBLE [30] feature set, (2) we compute the substitution cost between phones as the Hamming distance between their feature vector representations, and (3) we set the deletion and insertion cost to 0.5 which is roughly equivalent to the maximum possible substitution cost.

4. Evaluation Tasks

4.1. Acoustic Word Discrimination

The word discrimination task mainly evaluates the ability of a model to determine whether or not two given speech segments correspond to the same word type. We define this task as a segment-level retrieval task [31]: given a query segment $\overline{\boldsymbol{q}}$ and a candidate set of k word segments $\mathcal{S} = \{\overline{\boldsymbol{s}}_1, ..., \overline{\boldsymbol{s}}_k\}$, the goal is to rank segments in $\mathcal{S}$ in such a way that those segments corresponding to the same word type as the query $\overline{\boldsymbol{q}}$ are highly ranked. To this end, a vector-based search index is built by mapping each word segment in $\mathcal{S}$ into an embedding. Then, the cosine similarity between the embedding of the query $\overline{\boldsymbol{q}}$ and each embedding in the search index is computed which yields a ranked list, or an ordering $\mathcal{R}_{\boldsymbol{\theta}}$, of segments based on the cosine similarity score. The average precision metric is used to evaluate the quality of the ordering for a single query as

$$\text{AP} = \frac{1}{|\mathcal{S}_{\boldsymbol{q}}|} \sum_{r=1}^{k} P_{\boldsymbol{q}}(r) \times \mathcal{I}_{\boldsymbol{q}}(r) \tag{4}$$

where $\mathcal{S}_{\boldsymbol{q}}$ are the segments in $\mathcal{S}$ that correspond to the query $\overline{\boldsymbol{q}}$, $P_{\boldsymbol{q}}(r)$ is the precision at rank r, and $\mathcal{I}_{\boldsymbol{q}}(r)$ is a relevance function such that $\mathcal{I}_{\boldsymbol{q}}(r) = 1$ if the segment at rank r corresponds to the same word type as the query, or $\mathcal{I}_{\boldsymbol{q}}(r) = 0$ otherwise. The arithmetic average over all AP values in the test set yields the mean average precision (mAP) metric.

4.2. Word Phonological Similarity

To assess whether the emerging AWE space captures word-form similarity, we propose the word phonological similarity task. We argue that an evaluation based on phonological similarity will be more insightful to understand the impact of the model architecture and learning objective on the emergent embedding space. This evaluation task works as follows: given a query segment $\overline{\boldsymbol{q}}$ and a search index over the candidate set $\mathcal{S}$, two ranked lists, or orderings, are produced: (1) an ordering $\mathcal{R}_{\boldsymbol{\theta}}$ based on the cosine similarity between the AWEs, and (2) an ordering $\mathcal{R}_{\boldsymbol{\phi}}$ based on phonological similarity with the PWLD introduced in §3. To measure the degree of agreement between the two orderings $\mathcal{R}_{\boldsymbol{\theta}}$ and $\mathcal{R}_{\boldsymbol{\phi}}$, we use Kendall's τ, which is a measure of rank correlation between two ordinal variables [32]. For each query segment $\overline{\boldsymbol{q}}$ in the test set, Kendall's τ is computed as

$$\tau = 1 - \frac{2 \times \delta(\mathcal{R}_{\boldsymbol{\theta}}, \mathcal{R}_{\boldsymbol{\phi}})}{0.5 \times k(k-1)} \tag{5}$$

where $\delta(\mathcal{R}_{\boldsymbol{\theta}}, \mathcal{R}_{\boldsymbol{\phi}})$ is the minimum number of adjacent transpositions needed to bring $\mathcal{R}_{\boldsymbol{\theta}}$ to $\mathcal{R}_{\boldsymbol{\phi}}$. Kendall's τ coefficient takes values between 1.0 (identical ranks) and -1.0 (reverse ranks), while 0 indicates no association between the two orderings. Note that Spearman's correlation is not an appropriate metric for this task, as opposed to Kendall's τ which is designed to handle tied rankings that occur when the PWLD gives the same phonological distance for acoustic segments corresponding to the same word type.

Table 2: *Word-level statistics of our experimental data.*

	#segments per split			#phones ($\mu \pm$ std)	duration ($\mu \pm$ std)
	train	valid	test		
German	45886	7452	9964	6.8 ± 2.2	$.46 \pm 0.2$
Czech	68596	9244	11626	6.7 ± 2.5	$.50 \pm 0.2$

5. Experiments

5.1. Experimental Data

The data in our study is drawn from the GlobalPhone speech database for German and Czech [33]. We choose these two languages due to their predictable grapheme-to-phoneme (G2P) mapping and the availability of high quality G2P tools. We use the Montreal Force Aligner [34] to obtain time-aligned spoken word segments. Each acoustic word segment is parametrized as a sequence of 39-dimensional Mel-frequency spectral coefficients (MFSCs) where frames are extracted over intervals of 25ms with 10ms overlap. Table 2 shows summary statistics of our experimental data.

5.2. Architectures and Hyperparameters

CNN Acoustic Encoder. We employ a 3-layer temporal convolutional network (1D-CNN) with 256, 512, and 1024 filters and widths of 4, 8, and 16 for each layer and keep stride step at 1. Following each convolutional operation, we apply batch normalization, ReLU non-linearity, and dropout. We apply average pooling to downsample the representation at the end of the convolution block, which yields a 1024-dimensional AWE.

RNN Acoustic Encoder. We employ a 2-layer bidirectional Gated Recurrent Unit (BGRU) with a hidden state dimension of 512, which yields a 1024-dimensional AWE. We apply layer-wise dropout with a probability tuned over $\{0.0, 0.2, 0.4\}$.

Training Details. All models in this study are trained for 100 epochs with a batch size of 256 using the ADAM optimizer [35] and an initial learning rate (LR) of 0.001. The LR is reduced by a factor of 0.5 if the mAP on the validation set does not improve for 10 epochs. The epoch with the best validation performance during training is used for evaluation on the test set.

Implementation. We build our models using PyTorch [36] and use FAISS [37] for efficient similarity search. Our code is publicly available on GitHub[1].

[1] https://github.com/uds-lsv/AWEs_phon_sim

Table 3: *The results of our experiments for both tasks: word discrimination (mAP) and phonological similarity ($\overline{\tau}$).*

Encoder Architecture	Learning Objective	GERMAN		CZECH	
		mAP $\pm$ std	$\overline{\tau} \pm$ std	mAP $\pm$ std	$\overline{\tau} \pm$ std
Convolutional (1D-CNN)	PHONEDETECT	0.552 ± 0.33	0.043 ± 0.17	0.669 ± 0.31	0.080 ± 0.14
	WORD2PHONES	0.602 ± 0.33	0.127 ± 0.19	0.692 ± 0.31	0.150 ± 0.16
	SIAMESE (W/ RAND NEG)	0.621 ± 0.33	0.126 ± 0.12	0.731 ± 0.30	0.176 ± 0.10
	SIAMESE (W/ HARD NEG)	0.719 ± 0.32	0.074 ± 0.08	0.823 ± 0.27	0.093 ± 0.06
Recurrent (BGRU)	PHONEDETECT	0.652 ± 0.31	$\mathbf{0.237 \pm 0.14}$	0.730 ± 0.31	0.203 ± 0.11
	WORD2PHONES	0.692 ± 0.32	0.181 ± 0.15	0.796 ± 0.28	$\mathbf{0.226 \pm 0.13}$
	SIAMESE (W/ RAND NEG)	0.668 ± 0.34	0.148 ± 0.08	0.748 ± 0.31	0.153 ± 0.06
	SIAMESE (W/ HARD NEG)	$\mathbf{0.757 \pm 0.32}$	0.044 ± 0.05	$\mathbf{0.842 \pm 0.27}$	0.077 ± 0.04

5.3. Experimental Results

Our results are summarized in Table 3 for both evaluation tasks. The word discrimination task is measured by the mAP metric while the word phonological similarity task is measured by the mean of Kendall's τ rank correlation coefficients ($\overline{\tau}$).

Acoustic Word Discrimination. From the mAP values reported in Table 3, one can make two high-level observations: (1) all models outperform our classifier-based PHONEDETECT baseline, which is trained to detect phone n-grams in the acoustic segment, for both languages and architectures, and (2) recurrent models outperform their convolutional counterparts, which is consistent with the findings reported in the literature [10]. However, we also observe that the performance of the SIAMESE models, which are explicitly trained to minimize the cosine distance between segments of the same word type, largely depends on the negative sampling strategy. For example, in the case of recurrent SIAMESE models, we observe a relative improvement in the mAP score up to 13.32% for German and 12.57% for Czech when applying semi-hard negative sampling. Note that the SIAMESE recurrent models did not outperform the WORD2PHONES recurrent models when the contrastive negative samples were chosen randomly from the mini-batch. These findings highlight the importance of negative sampling in training AWEs with contrastive objectives, which is a matter that has not been previously investigated to the best of our knowledge. We conclude that models trained with objectives that explicitly optimize the distance in the AWE space outperform other models that lack this objective on word discrimination, especially if the negative samples are chosen with a challenging criterion.

Word Phonological Similarity. In this evaluation, we observe a positive correlation between the distance in the embedding space and phonological distance for all models as indicated by the positive values of the $\overline{\tau}$ metric. Nevertheless, the correlation seems to be either weak or moderate in the best cases. Moreover, we observe that both the learning objective and encoder architecture have a considerable impact on the extent to which the embedding space captures phonological similarity. For example, the convolutional PHONEDETECT models show some of the lowest correlation scores (German $\overline{\tau}$ = 0.043 and Czech $\overline{\tau}$ = 0.080), while their recurrent counterparts show some of the highest correlation scores (German $\overline{\tau}$ = 0.237 and Czech $\overline{\tau}$ = 0.203), despite having the same training objective. These findings indicate that convolutional encoders may behave like shallow pattern detectors when trained as a classifier while recurrent encoders tend to preserve the temporal structure of the acoustic input in their representation. Another observation that we find surprising is the poor performance of the SIAMESE models on this task given that they outperform the other models on word discrimination. Overall, recurrent models which are trained with symbolic grounding (namely PHONEDETECT and WORD2PHONES) are better at capturing word-form similarity compared to their convolutional counterparts on the one hand, and the recurrent SIAMESE models on the other.

6. Discussion and Conclusion

Although the vast majority of previous work has been driven by the engineering applications of AWEs, there is a growing scientific interest in using deep neural networks as cognitive models of (human) speech processing [38, 22, 23, 39, 40]. Therefore, we argue that this cognitively motivated direction requires us to take a closer look at the embedding space and examine the degree to which we can rely on the emergent distance as an estimate of (perceptual) dissimilarity between linguistic units. In this paper, we take a step in this direction and conduct a set of experiments where we keep the training conditions for each model fixed and systematically study the impact of the architecture (convolutional and recurrent) and learning objective (classification, phonological decoding, and contrastive objectives) on the AWEs' performance using two evaluation tasks: acoustic word discrimination and word phonological similarity.

Our experiments demonstrate that while contrastive objectives yield AWEs with strong discriminative performance, they fail to reflect the phonological distance between word-forms, especially compared to AWEs which are trained with symbolic grounding (i.e., phone sequences corresponding to words). We hypothesize that the contrastive objective emphasizes word separability in the embedding space which hinders the ability of the emerging distance to reflect word similarity. Moreover, our experiments show a consistent trend with recurrent models outperforming their convolutional counterparts in both evaluation tasks, which we attribute to the ability of recurrent DNNs to model the temporal nature of speech. In conclusion, our experimental findings highlight the necessity for more diverse evaluation schemes when working with AWEs to investigate the degree to which they produce human-like errors. Furthermore, our work can be extended by analyzing the correlation between embedding distances and human perceptual similarity judgments.

7. Acknowledgements

We thank the anonymous reviewers for their constructive comments. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 232722074 – SFB 1102.

8. References

[1] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Proc. Interspeech*, 2010.

[2] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2011.

[3] A. Anastasopoulos, S. Bansal, D. Chiang, S. Goldwater, and A. Lopez, "Spoken term discovery for language documentation using translations," in *Proc. of the Workshop on Speech-Centric Natural Language Processing*, 2017.

[4] 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.

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

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

[7] 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.

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

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

[12] G. Heigold, P. Nguyen, M. Weintraub, and V. Vanhoucke, "Investigations on exemplar-based features for speech recognition towards thousands of hours of unsupervised, noisy data," in *Proc. ICASSP*, 2012.

[13] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," in *Proc. ICASSP*, 2007.

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

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

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

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

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

[19] 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.

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

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

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

[23] 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.

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

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

[26] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Proc. NIPS*, 1994.

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

[28] 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.

[29] L. Fontan, I. Ferrané, J. Farinas, J. Pinquier, and X. Aumont, "Using phonologically weighted levenshtein distances for the prediction of microscopic intelligibility," in *Proc. Interspeech*, 2016.

[30] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/

[31] M. Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

[32] M. Lapata, "Automatic evaluation of information ordering: Kendall's Tau," *Computational Linguistics*, 2006.

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

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

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

[36] 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.

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

[38] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, 2018.

[39] J. S. Magnuson, H. You, S. Luthra, M. Li, H. Nam, M. Escabi, K. Brown, P. D. Allopenna, R. M. Theodore, N. Monto *et al.*, "Earshot: A minimal neural network model of incremental human speech recognition," *Cognitive science*, vol. 44, no. 4, p. e12823, 2020.

[40] 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.