ANALYZING AUTOENCODER-BASED ACOUSTIC WORD EMBEDDINGS

Yevgen Matusevych
School of Informatics
University of Edinburgh
ymatusev@ed.ac.uk

Herman Kamper
E&E Engineering
Stellenbosch University
kamperh@sun.ac.za

Sharon Goldwater
School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

ABSTRACT

Recent studies have introduced methods for learning acoustic word embeddings (AWEs)—fixed-size vector representations of words which encode their acoustic features. Despite the widespread use of AWEs in speech processing research, they have only been evaluated quantitatively in their ability to discriminate between whole word tokens. To better understand the applications of AWEs in various downstream tasks and in cognitive modeling, we need to analyze the representation spaces of AWEs. Here we analyze basic properties of AWE spaces learned by a sequence-to-sequence encoder-decoder model in six typologically diverse languages. We first show that these AWEs preserve some information about words' absolute duration and speaker. At the same time, the representation space of these AWEs is organized such that the distance between words' embeddings increases with those words' phonetic dissimilarity. Finally, the AWEs exhibit a word onset bias, similar to patterns reported in various studies on human speech processing and lexical access. We argue this is a promising result and encourage further evaluation of AWEs as a potentially useful tool in cognitive science, which could provide a link between speech processing and lexical memory.

1 INTRODUCTION

Several recent studies have introduced acoustic word embeddings (AWEs). AWEs are vector representations of individual word tokens based on their acoustic features (Levin et al., 2013; Chung et al., 2016; Holzenberger et al., 2018; Kamper, 2019, etc.) rather than on their relation to other words, as in semantic (textual) word embeddings.[1] Acoustic words unfold dynamically in time and have variable duration, yet fixed-dimensional AWEs have shown good performance in speech processing tasks such as word discrimination, and a recent study also suggests they can correctly predict some patterns of infant phonetic learning (Matusevych et al., 2020). These results encourage exploration of AWEs for cognitive modeling, just as semantic word embeddings have been used as models of human semantic memory (e.g., Grand et al., 2018; Nematzadeh et al., 2017; Pereira et al., 2016). As a first step, we need to describe the basic properties of AWEs and compare them to patterns observed in human lexical memory and spoken word perception, to better understand how temporal sequences of phones are encoded into static holistic representations, and whether these representations might correspond to human lexical representations. To our knowledge, one existing study (Ghannay et al., 2016) evaluates properties of AWEs, but it focuses on comparing them to orthographic word embeddings and only considers one language, French.

In this study, we consider AWEs in six different languages generated by a recent speech representation learning model, a correspondence-autoencoding recurrent neural network (CAE-RNN; Kamper, 2019), and analyze their basic properties to understand the organizing principles of the AWE space. An acoustic word contains three types of signal: (i) properties specific to the particular instance of the word (in this study, we focus on one such feature, absolute duration), (ii) the speakers' characteristics (i.e., all acoustic words spoken by the same person share some acoustic properties), and (iii) the word's phonetic properties (i.e., *cat* is more similar to *catch* than to *dog*). Like many other AWE

---

[1] Although see Chung & Glass (2018); Chung et al. (2018) for speech-based *semantic* embeddings, which we do not consider here.

models, the CAE-RNN is designed to abstract away from the first two types of information and learn the similarities between various spoken instances of the same word, similar to spoken word recognition in human speakers, who can identify the wordform (lexical item) regardless of who pronounces it and how. Existing work shows that the CAE-RNN succeeds in doing this: relative to a baseline that uses traditional signal processing methods, it is better at discriminating between pairs of same vs. different words and at clustering together different instances of the same word in its embedding space (Kamper, 2019; Kamper et al., 2020). At the same time, we show here that AWEs generated by the CAE-RNN do not completely abstract away from the information about an acoustic word's absolute duration and speaker identity. More interestingly from a cognitive perspective, we also demonstrate that the AWEs exhibit a word onset bias, corresponding to a broad range of patterns reported in literature on human speech processing and lexical access which suggest that humans consider the initial sound of the word more 'prominent' than its other sounds: for example, speakers emphasize it in articulation (Fougeron & Keating, 1997; Keating et al., 1999), listeners can capture the distinctions between word-initial and word-final sounds (Shatzman & McQueen, 2006), initial sounds have a special status in spoken word recognition (Marslen-Wilson & Zwitserlood, 1989), and the first letter is a more efficient cue for lexical retrieval than other letters (Brown & Knight, 1990).

2 METHOD

The CAE-RNN model (Kamper, 2019), which we use for obtaining AWEs, is inspired by a sequence-to-sequence autoencoder, in which both the encoder and the decoder are RNNs (Chung et al., 2016). During training, the CAE-RNN receives two different instances of the same wordform at a time: it encodes one of them into a vector of fixed dimensionality and uses this vector to reconstruct the other instance sequentially. Each instance is represented as a sequence of *frames*, where a frame is a 13-dimensional vector of mel-frequency cepstral coefficients (a standard way of representing the energy spectrum) extracted from a 25-ms-long slice of speech. Learning the correspondence between two instances of the same word encourages the model to abstract away from random noise and speaker characteristics while learning to encode the word-invariant phonetic information. This top-down guidance from the word level finds parallels in studies showing that even 6–8-month infants can recognize some wordforms in running speech (e.g., Jusczyk & Aslin, 1995; Jusczyk et al., 1999), and that this information can be useful for learning phonetic information (Feldman et al., 2013).

Following Kamper et al. (2020), we train a set of models on six typologically diverse languages (see Appendix for details) from the GlobalPhone corpus (Schultz, 2002). Using the encoder of each model, we obtain AWEs for a set of unseen test words in the corresponding language. On these AWEs, we run a series of tests focusing on three main questions: (1) Do these AWEs preserve some information about speaker characteristics and segment acoustic properties (namely, its duration)? (2) Can AWEs abstract away from these two types of signal in favor of linguistically meaningful information, such as word phonetic similarity? (3) Can AWEs exhibit the human-like word onset bias? To address these questions, we probe the structure of the AWEs using three methods: (i) using linear classifiers[2] or regressions trained on top of AWEs; (ii) using a machine ABX task (Schatz et al., 2013), in which the distance between words A and X is compared to the distance between words B and X; and (iii) directly comparing the distances between pairs of words meeting specific criteria.

To see how much the results rely on representation learning, we compare to a simple *downsampling* baseline (DS; Holzenberger et al., 2018), which creates 130-dimensional embeddings (the same as the CAE-RNN AWEs) by concatenating 10 frames from the input word, equally spaced in time.

3 EXPERIMENTS AND RESULTS

Speaker identity. First, we look at whether the AWEs preserve any information about speaker identity, despite being trained to ignore the variation across speakers. We train a multiclass logistic regression classifier on 80% of the embedded words to predict the speaker identity, and then test it on the held-out 20% of words. Figure 1 shows that the learned AWEs predict speaker identity worse than the DS baseline, but better than the majority class baseline: that is, they abstract away from speaker characteristics to some degree, but not completely.

---

[2]Linear classifiers allow for making claims about linear separability of the classes in an embedding space, a finding much easier to interpret than a potentially high performance of a complex nonlinear classifier.
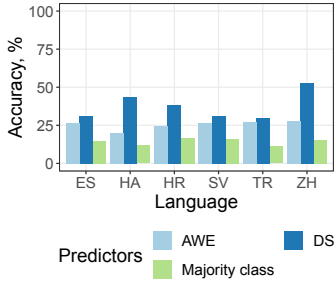
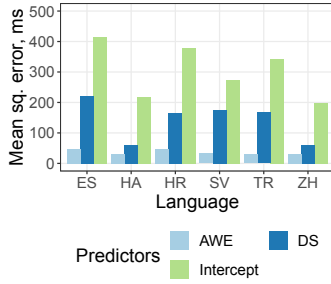Figure 1: Classification accuracy of models predicting a word's speaker identity.

Figure 2: Mean squared error of linear models predicting absolute word duration.
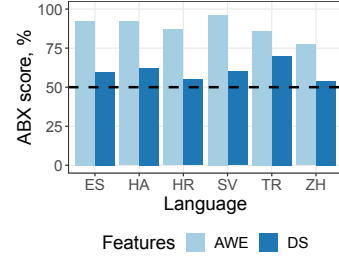
Figure 3: ABX scores in the task with words A and X matched on word duration, and B and X on speaker identity.

Word duration. Next, we test whether the fixed-dimensional AWEs preserve information about a basic acoustic property of a word—its absolute duration in milliseconds—without such information being explicitly provided. For each language, we train a linear regression model on 80% of the embedded words to predict the word's absolute duration, and then test the model on the held-out 20% of the words. Figure 2 shows that the learned AWEs predict the word duration better than the DS baseline and the intercept baseline (i.e., a linear regression that only fits an intercept, thus always predicting the mean duration), with R^2 in the range 0.85–0.91 (not shown in Figure 2). This suggests that the AWEs successfully encode temporal sequences into fixed-dimensional vectors while preserving information about their length. However, a word's absolute duration reflects not only random variation in the speech rate (i.e., duration as an acoustic property of the word as a speech segment), but also the number of phones in the word (i.e., the word's phonetic properties): *category* on average takes longer to say than *cat*. To consider the duration as a purely acoustic property, we next look at various instances of the same word.

Segment duration vs. speaker identity. We know that our AWEs encode some information about both segment duration and speaker identity, but do they encode both kinds of signal equally well? To test this, we design an ABX task with three instances of the same word, where A and X are generated by different speakers (but have similar duration, within a factor of 1.1), while B and X are generated by the same speaker (but are different in duration by a factor of at least 1.5). A score higher than 50% indicates that word duration is encoded to a higher degree in the embedding space, while a score lower than 50% shows that speaker identity is encoded better. Figure 3 shows that, while the DS baseline performs nearly at chance for 4 out of 6 languages, in our AWEs the absolute duration is a more distinctive feature than the speaker identity. Note that in this case there are no *phonetic* differences between the acoustic words, which suggests the segment's duration is encoded in the AWEs as an *acoustic* feature. Next, we test whether the AWEs also encode phonetic information.

Number of phones. To see how well our AWEs encode linguistically meaningful information, we look at the properties related to the words' phonetic content. First, we test whether the AWEs encode the information about the number of phones in a word. We train/test a linear regression model to predict the number of phones in the words, using an $80/20\%$ split, as before. Figure 4 shows that the AWEs predict the number of phones better than both the DS data and the intercept baseline (i.e., a linear regression always predicting the mean value), with R^2 in the range 0.71–0.84 (not shown in Figure 4), suggesting that the AWE encode some information about the number of phones.

Phonetic similarity. If our AWEs also encode words' phonetic properties, we expect phonetically similar words to be closer in the embedding space than dissimilar words. To test this, we look at whether the cosine distance between pairs of AWEs increases with the phone edit distance between the words (i.e., phonetic dissimilarity). Figure 5 shows the results for Hausa (the trends are similar in the other languages): we observe the expected trend both in the DS baseline and in the AWEs, but in the AWEs words that are more phonetically similar have more similar representations compared to the DS (which is especially evident for the pairs with edit distance zero: instances of the same word and/or homophones). This confirms that our AWEs encode some of the words' phonetic properties.

Word onset bias. Finally, we ask whether the AWEs exhibit the human-like word onset bias: considering the first sound of the word more 'prominent' than its other sounds. We use an ABX task
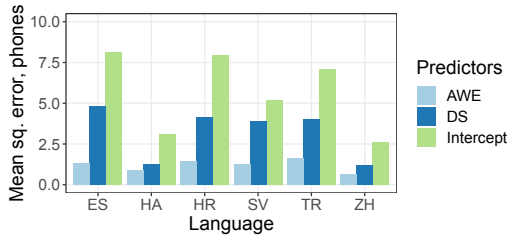
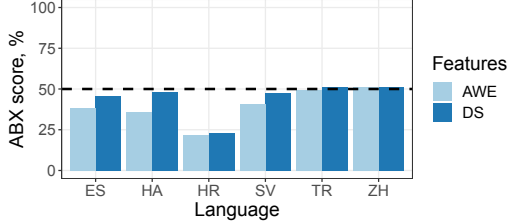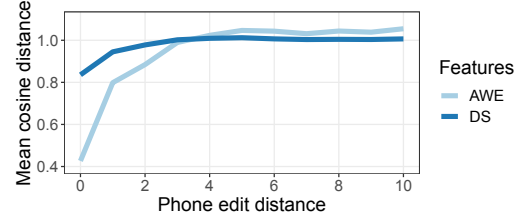Figure 4: Mean squared error of linear models predicting number of phones in a word.

Figure 5: Phone edit distance between pairs of Hausa acoustic words against average cosine distance between their representations.

Figure 6: ABX scores in the task where words A and X differ in their initial phone, and B and X in another phone.
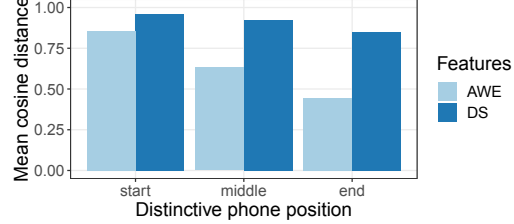

Figure 7: Average cosine distances between pairs of Hausa AWEs for words differing in one phone, depending on the position of that phone.

and a comparison of distances between words, in both methods focusing on pairs of words with phone edit distance of 1. In the ABX task, words A and X differ in their first phone, while B and X differ in another phone (e.g., X: *take*, A: *cake*, B: *ta**p**e*). A score of 50% indicates no difference depending on the distinctive phone position (i.e., X is equally close to A and B), a score below 50% indicates the expected bias (i.e., X is closer to A than to B), and a score above 50% indicates a bias in the opposite direction. Figure 6 shows that the AWEs score below 50% in most languages, indicating a larger distance between words that differ in their first phone compared to words that differ in another phone, which corresponds to the predicted bias. Importantly, the scores are lower in the AWEs than in the DS data, suggesting that this bias does not completely arise from the data alone, but is learned by the model (although the presence of the bias in the DS data suggests that the first phone may provide a stronger signal—e.g., in terms of duration—than other phones). In addition, when we look at the distances between pairs of AWEs for words that differ in a single phone in Hausa (Figure 7, with similar results in other languages), we observe larger distances when the distinctive phone is at the beginning of the word (rather than in the middle or at the end), and this tendency is stronger in the AWEs than in the DS data, in line with the results of our ABX task.

4 CONCLUSION

We presented an analysis of basic properties of a particular type of acoustic word embeddings, which are based on an encoder-decoder model. We showed that these embeddings can succeed in encoding some characteristics of the words' phonetic content, yet they also preserve information about an acoustic word's absolute duration and speaker identity. We also show that AWEs can exhibit a bias towards treating the first sound in the word as a more important part of the signal, compared to the other sounds—a pattern mirroring the empirical data observed in human speakers. These results suggest that AWEs show some promise as a modeling tool in cognitive science, and encourage further research in this direction. AWEs can provide a straightforward connection between human speech processing and lexical storage and access, as acoustic words of any duration are situated within a feature space that is easy to probe with various tests such as the ones presented in this study. While AWEs are devoid of any semantics, they could be combined with speech-based (Chung & Glass, 2018) or textual semantic word embeddings (as in Chen et al., 2018), potentially informing more accurate models of human lexical memory and access, which need to take into account word pronunciations or their acoustic properties (Aydelott & Bates, 2004; Andruski et al., 1994).

ACKNOWLEDGEMENTS

REFERENCES

Jean E. Andruski, Sheila E. Blumstein, and Martha Burton. The effect of subphonetic differences on lexical access. *Cognition*, 52:163–187, 1994.

Jennifer Aydelott and Elizabeth Bates. Effects of acoustic distortion and semantic context on lexical access. *Language and Cognitive Processes*, 19:29–56, 2004.

Alan S. Brown and Kevin K. Knight. Letter cues as retrieval aids in semantic memory. *The American Journal of Psychology*, pp. 101–113, 1990.

Yi-Chen Chen, Sung-Feng Huang, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 941–948, 2018.

Yu-An Chung and James Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Proceedings of Interspeech*, pp. 811–815, 2018.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Proceedings of Interspeech*, pp. 765–769, 2016.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems*, pp. 7354–7364, 2018.

Naomi H. Feldman, Thomas L. Griffiths, Sharon Goldwater, and James L. Morgan. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120:751–778, 2013.

Cécile Fougeron and Patricia A. Keating. Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, 101:3728–3740, 1997.

Sahar Ghannay, Yannick Estève, Nathalie Camelin, and Paul Deléglise. Evaluation of acoustic word embeddings. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 62–66, 2016.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection: Recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv:1802.01241*, 2018.

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux. Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments. In *Proceedings of Interspeech*, pp. 2683–2687, 2018.

Peter W. Jusczyk and Richard N. Aslin. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29:1–23, 1995.

Peter W. Jusczyk, Derek M. Houston, and Mary Newsome. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39:159–207, 1999.

Herman Kamper. Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In *Proceedings of ICASSP*, pp. 6535–3539, 2019.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. Multilingual acoustic word embedding models for processing zero-resource languages. *arXiv:2002.02109*, 2020.

Patricia Keating, Richard Wright, and Jie Zhang. Word-level asymmetries in consonant articulation. *UCLA Working Papers in Phonetics*, pp. 157–173, 1999.

Keith Levin, Katharine Henry, Aren Jansen, and Karen Livescu. Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 410–415, 2013.

William Marslen-Wilson and Pienie Zwitserlood. Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15:576–585, 1989.

Yevgen Matusevych, Thomas Schatz, Herman Kamper, Naomi Feldman, and Sharon Goldwater. Evaluating computational models of infant phonetic learning across languages. *Under review*, 2020.

Aida Nematzadeh, Stephan C. Meylan, and Thomas L. Griffiths. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of CogSci*, pp. 859–864, 2017.

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, 33:175–190, 2016.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proceedings of Interspeech*, pp. 1781–1785, 2013.

Tanja Schultz. GlobalPhone: A multilingual speech and text database developed at Karlsruhe University. In *Proceedings of ICSLP*, pp. 345–348, 2002.

Keren B. Shatzman and James M. McQueen. Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, 68:1–16, 2006.

A APPENDIX

Model training. We train six monolingual CAE-RNN models (one per language) on data extracted from GlobalPhone, a non-parallel corpus of read newspaper articles (Schultz, 2002). Each language has 16 hours of training and 2 hours of test data on average, with test data sampled from held-out speakers. To prepare word pairs for training the model, we first create a list of all words in the training data (obtained through forced word alignments) with duration of at least 500 ms and containing at least 5 phones. We then randomly pair words of the same type to create $100,000$ pairs. Following Kamper et al. (2020), we pre-train the model as an autoencoder for 15 epochs to initialize its parameters, and then train it for 25 epochs using the existing architecture: 3 hidden layers (400 gated recurrent units each) in both the decoder and encoder, and an embedding dimensionality of 130. For reference, on the 'same-different' task, these models score 60–85%, depending on the language.

Code	Language	Family (branch)	Test speakers	Phones per word: mean (SD)
ES	Spanish	Indo-European (Romance)	10	4.9 (3.1)
HA	Hausa	Afroasiatic (Chadic)	10	4.2 (1.8)
HR	Croatian	Indo-European (Slavic)	10	5.4 (2.8)
SV	Swedish	Indo-European (Germanic)	9	4.1 (2.3)
TR	Turkish	Turkic (Oghuz)	11	6.0 (2.7)
ZH	Mandarin	Sino-Tibetan (Mandarin)	11	3.6 (1.6)

Table 1: Characteristics of the test data.