12-2022

Supervised Acoustic Embeddings And Their Transferability Across Languages

Sreepratha Ram
United Arab Emirates University

Hanan Aldarmaki
Mohamed bin Zayed University of Artificial Intelligence

Recommended Citation

S. Ram and H. Aldarmaki, "Supervised Acoustic Embeddings And Their Transferability Across Languages", In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022),* ACL, pp. 212–218, Dec 2022. https://aclanthology.org/2022.icnlsp-1.24

Supervised Acoustic Embeddings And Their Transferability Across Languages

Sreepratha Ram
UAE University
sree_ram@uaeu.ac.ae

Hanan Aldarmaki
MBZUAI
hanan.aldarmaki@mbzuai.ac.ae

Abstract

In speech recognition, it is essential to model the phonetic content of the input signal while discarding irrelevant factors such as speaker variations and noise, which is challenging in low-resource settings. Self-supervised pre-training has been proposed as a way to improve both supervised and unsupervised speech recognition, including frame-level feature representations and Acoustic Word Embeddings (AWE) for variable-length segments. However, self-supervised models alone cannot learn perfect separation of the linguistic content as they are trained to optimize indirect objectives. In this work, we experiment with different pre-trained self-supervised features as input to AWE models and show that they work best within a supervised framework. Models trained on English can be transferred to other languages with no adaptation and outperform self-supervised models trained solely on the target languages.

***Keywords*—** Unsupervised ASR, Transfer Learning, Acoustic Word Embeddings

1 Introduction

With supervised speech recognition systems getting more robust and accurate due to the availability of large amounts of labeled data and computational power (Gulati et al., 2020; Baevski et al., 2020b), more attention is now given to low-resource languages for which training data are limited or non-existent (Aldarmaki et al., 2022). Unsupervised pre-training using unlabeled speech can be leveraged to improve both supervised and unsupervised models; for instance, speech representations pre-trained on large amounts of unlabeled speech from multiple languages have been shown to improve ASR performance for low-resource languages (Kawakami et al., 2020; Conneau et al., 2020).

While most supervised ASR models operate at the level of phones, word-level segmental ASR where variable-length segments are modeled and embedded into fixed-dimensional vectors have also been explored with relative success (Abdel-Hamid et al., 2013; He and Fosler-Lussier, 2015). In a similar vein, Acoustic Word Embeddings (AWEs) have been proposed as a way to efficiently compare variable-length speech segments in low-resource settings (Peng et al., 2020; Kamper et al., 2020). Unlike written words, spoken words naturally contain speaker and phonetic variability that makes them more difficult to model in a latent space without supervision. Self-supervised pre-training and cross-lingual transfer are two possible approaches to make unsupervised models more robust to non-linguistic variations in the input signal.

In this work, we investigate the performance of self-supervised training of AWE models versus supervised training with zero-shot cross-lingual transfer. We experiment with different types of acoustic features and measure their performance separately and within the AWE models. While we find that pre-trained acoustic features improve the performance of self-supervised AWE models to some extent, a larger improvement can be achieved when the AWE models are trained in a supervised manner using small amount of labeled data from a different language. This zero-shot cross-lingual transfer is observed consistently across different languages, and particularly with the use of pre-trained feature representations. Our results suggest that supervised training with zero-shot cross-lingual transfer is a more effective approach for low-resource speech models compared with purely self-supervised training[1].

2 Background & Related Work

Spoken language is often modeled using short fixed-length frames of 10 to 30 ms duration, which

[1]We provide python training and evaluation scripts for replicating our experiments: https://github.com/h-aldarmaki/acoustic_embeddings

results in variable-length word segments. Dynamic Time Warping (DTW) is an early technique that uses dynamic programming to compare variable-length segments by finding optimal frame-wise alignment. DTW is rather inefficient, which motivates embedding variable-length segments into vectors of fixed size that can be compared using more efficient metrics such as cosine or Euclidean distance (Levin et al., 2013). Different types of Acoustic Word Embeddings (AWE) have been proposed. As these techniques are generally meant for low-resource languages, they are typically trained in a self-supervised manner, most commonly using an auto-encoder network with reconstruction loss (Chung et al., 2016; Holzenberger et al., 2018). Compared with direct comparison via DTW, these AWEs generally result in similar or slightly superior performance while being far more efficient (Holzenberger et al., 2018). Peng et al. (2020) describes an alternative training strategy using correspondence auto-encoders, which relies on word pairs extracted via unsupervised spoken term discovery, and further improvements can be achieved using contrastive learning and multi-lingual adaptation (Jacobs et al., 2021).

The above models use static acoustic features (e.g. MFCCs) as input. van Staden and Kamper (2021) shows that using pre-trained features like CPC (van den Oord et al., 2018) improves the performance of unsupervised AWE models. Pre-trained features have been repeatedly shown to improve performance in supervised downstream tasks (Yang et al., 2021). In addition, pre-trained features have been shown to transfer across languages. For instance, a modified version of CPC (MCPC) is described in Riviere et al. (2020), which demonstrates that pre-training these features on Egnlish results in improved phone classification accuracy for other languages. Other types of pre-trained features, such as wav2vec 2.0 (Baevski et al., 2020a) have been shown to improve both supervised and unsupervised ASR performance (Baevski et al., 2021), and multi-lingual training of these features (i.e. XLSR-53) can lead to improvements across many languages compared to monolingual pre-training (Conneau et al., 2020).

3 Objectives & Methodology

The objective of this study is to investigate the effectiveness and trasnsferability of pre-trained acoustic features when used as input to acoustic word embeddings. To that end, we compare self-supervised AWEs trained directly on the target languages versus zero-shot cross-lingual transfer of supervised AWEs trained on a different source language. To our knowledge, the combination of pre-trained features with AWE models has not been fully investigated; most AWE models are trained with standard acoustic features like MFCCs, while self-supervised features are typically evaluated within supervised models fine-tuned for the target languages. Furthermore, zero-shot cross-lingual transfer of supervised AWEs has not been the focus of previous works in this area, which mainly focused on improving self-supervised AWEs.

For the purpose of this evaluation, we use a relatively simple architecture for the embedding model and we fix the hyper-parameters based on preliminary validation results for English self-supervised AWEs[2]. We do not do any further tuning of the self-supervised or the supervised models. We use English as the source language, and evaluate zero-shot transfer on four other languages: French, German, Spanish, and Arabic, with the latter used as a challenge set since it contains more variability and noise. No labeled data were used for the target languages with the exception of word boundaries which were obtained via force alignment. We evaluate mainly using minimal-pair ABX error rates to measure phonetic discriminability and speaker invariance. We also cluster the embedded words and measure how often different occurrences of the same words end up in the same cluster.

4 Experimental Settings

4.1 Model Architecture

Our AWE model consists of a multi-layer bidirectional LSTM encoder, followed by a uni-directional LSTM decoder, similar to Chung et al. (2016) and (Holzenberger et al., 2018). The encoder takes a sequence of T acoustic features representing one spoken word. The forward and backward states of the last hidden layer of the encoder are concatenated and used as an embedding of the given word, call it $\mathbf{h}^T$. The decoder generates the target sequence one step at a time, conditioned on $\mathbf{h}^T$ and the output at the previous time step, similar to Chung and

[2]We observed that self-supervised models were very sensitive to the choice of architecture and hyper-parameters, so we fixed these in favor of self-supervised models. As shown in later sections, we still got better results with the supervised models, which shows that they are more robust and easier to optimize on top of being more effective.

Glass (2018). In the self-supervised setting, the target sequence is the same as the input sequence, so the model is trained as an auto-encoder with MSE loss. In the supervised setting, the target is a sequence of phonemes representing the input word, and the model is trained by minimizing the negative log-likelihood. We used 2-layer networks with 100 hidden units for most models, which results in embeddings of size 200. We also used dropout with probability 0.3 on the input features, similar to the denoising networks used in Chung et al. (2016). More details of the parameters and training process can be found in the Appendix.

4.2 Feature Extraction

For easier reproduciblity, we used the s3prl toolkit[3] for extracting all features. We used the pre-trained s3prl upstream models; among the many pretrained self supervised speech representations available, modified CPC, Wav2Wec2 and XLSR-53 were chosen based on superior DTW-based ABX scores[4]. All pre-trained models, with the exception of XLSR, have been exclusively pre-trained on English data. XLSR-53 was pre-trained on unlabeled speech from 53 languages, including all target languages in our experiments. As observed by other researchers (Bartelds et al., 2022), the performance of features extracted from transformer-based models is largely dependent on the choice of layer; we used the last hidden layer for modified CPC, the second to last hidden layer for Wav2Vec2 and the central hidden layer (layer 12) for XLSR-53. Averaging all layers gave reasonable results, but these choices led to the best performance. For MFCC features, we also used the s3prl implementation, which includes 13 static features as well as dynamic delta and delta-delta coefficients.

4.3 Data

We used the Librispeech (Panayotov et al., 2015) and Multilingual Librispeech (Pratap et al., 2020) datasets for English (en), French (fr), German (de), and Spanish (es). We used the dev sets for training, and test sets for evaluation (dev-clean and test-clean for English). We obtained the word boundaries automatically by forced alignment. For Arabic (ar), we used the dev and test sets of MGB2 (Ali et al., 2016). This dataset is expected to be more challenging as it contains a diversity of dialects as well as various noise conditions. See the Appendix for more details on the datasets and the word alignment process.

4.4 Evaluation Scheme

We constructed Minimal-Pair ABX tasks, as described in Schatz et al. (2013). ABX tasks are typically used to measure phoneme discrimination in zero-resource settings, and they consist of two segments, A and B, that differ by a minimal contrast (e.g. one phoneme difference), and a third segment X that matches either A or B. A distance measure such as DTW or cosine is used to find the closest match. We used two variants of this task: within-speaker ABX, where all three words are spoken by the same speaker, and cross-speaker ABX, where X is spoken by a different speaker. We automatically extracted the words from each test set; we selected A and B by finding word pairs that have the same length[5] and Levenshtein edit distance of 1 or 2, which roughly corresponds to a difference of one or two phonemes most of the time. For Arabic, the dataset did not have speaker ids, so all three words could be from different speakers. In addition, due to the lower quality of the sound recordings and the presence of noise in this dataset, the word alignment quality is much lower than the other languages, so the automatic process resulted in many invalid segments. To have a more reliable test set for Arabic, we manually checked the validity of the extracted words and kept 954 validated word pairs for evaluation.

We also used clustering for complementary evaluation. We clustered the embeddings using K-Means with K being the number of unique words in the test set. We calculated the accuracy of clustering as the percentage of words that match their cluster label, which is the word id of the majority of segments in each cluster. This allows us to measure if the embeddings of the same words are similar enough to be clustered together.

5 Results

Table 1 shows ABX error rates using the input features directly (with DTW as distance metric), self-supervised AWEs trained on each language,

[3] https://github.com/s3prl/s3prl

[4] We did experiment with other features like APC, VQ-APC, and VQ-Wav2Vec, and got similar or inferior performance to MCPC and Wav2Vec2. We opted to omit these for brevity.

[5] Since automatic word alignments tend to be inaccurate around the boundaries, we only used words that have at least five characters.

	en		fr		de		es		ar
	within	across	within	across	within	across	within	across	across
Using DTW									
MFCC	9.98	19.85	12.59	24.82	11.46	25.03	11.83	25.27	40.98
wav2vec	8.51	11.15	9.95	15.61	9.01	15.08	10.13	15.46	37.42
MCPC	7.80	11.74	9.96	17.09	9.22	15.85	11.14	18.03	38.99
XLSR-53	9.45	13.72	10.97	16.77	10.89	15.76	14.31	19.31	40.15
Self-Supervised AWEs in each language									
MFCC	12.30	19.12	16.63	25.06	16.71	25.99	16.58	25.21	43.92
wav2vec	6.63	9.27	9.94	13.19	10.56	15.09	12.25	15.23	38.16
MCPC	7.66	9.53	11.64	16.19	10.24	15.30	13.25	16.29	41.09
XLSR-53	10.61	12.19	13.72	16.19	12.59	15.73	17.10	20.14	37.00
Supervised AWEs trained on English									
MFCC	3.83	4.57	10.77	15.32	9.16	13.06	11.49	16.56	38.15
wav2vec	1.38	1.14	6.59	9.44	4.98	7.32	7.32	10.12	34.80
MCPC	2.49	2.51	8.13	12.23	6.66	10.68	9.89	13.99	39.20
XLSR-53	**0.93**	**0.79**	**4.12**	**5.71**	**1.92**	**2.83**	**5.05**	**6.21**	**31.76**

Table 1: ABX error rates (%) within and across speakers for each language

	en	fr	de	es	ar
Self-Supervised AWEs for each language					
MFCC	47.3	48.4	45.0	54.3	30.9
wav2vec	66.1	59.9	59.5	67.5	35.9
MCPC	57.2	53.6	53.9	60.4	33.5
XLSR-53	52.0	52.2	54.9	55.0	31.0
Supervised AWEs trained on English					
MFCC	68.5	52.7	56.7	61.1	33.4
wav2vec	82.3	64.8	69.3	71.1	38.8
MCPC	74.5	56.4	62.7	66.2	35.6
XLSR-53	**84.3**	**69.1**	**78.1**	**75.8**	**41.8**

Table 2: K-Means Clustering Accuracy (%)

and supervised AWEs trained on English. Cosine similarity is the metric used in the latter two settings. Confirming previous results (Riviere et al., 2020), we do observe that pre-trained acoustic features like Modified CPC and Wav2Vec2, which are trained exclusively on English unlabeled speech, transfer well across languages. These pre-trained features consistently outperformed MFCC features for all languages, particularly in cross-speaker evaluation. Unsurprisingly, the English language has the best ABX scores overall simply because the pre-trained features used are all trained on English. The results for self-supervised AWE models are mixed, but generally they are in the same range as DTW performance, which also conforms with previously published results (Holzenberger et al., 2018).

With supervised training, we see significant reduction in errors rates for all languages. The lowest error rates are achieved on the English test set, as expected. More notably, the largest reduction in error rates is achieved with the XLSR features. It is also interesting to note that XLSR features were not impressive in the self-supervised setting compared with other features; Wav2Vec2 and MCPC, which were trained on English only, gave better results in the self-supervised framework for all test languages. The advantage of using these cross-lingual features was only evident in the supervised and transfer learning setting, where they consistently outperformed all other features. For Arabic, the error rates are higher overall due to the nature of the dataset, but we still observe the lowest error rate in the transfer learning setting.

Finally, we see in table 2 that the clustering accuracy results are consistent with the ABX results, where supervised models trained on English consistently gave higher accuracy compared with self-supervised models trained on the target languages.

6 Conclusions

Our results demonstrate the superior effectiveness of zero-shot transfer learning of acoustic word embeddings compared with self-supervised training in the target languages. This is particularly useful for low-resource languages for which data may not be available for supervised or self-supervised

training. The mechanism of this transfer is mainly through the reduction in speaker variability which is far easier to achieve via supervised training. In addition, supervised training makes the most out of pre-trained features, where we see further reduction in error rates that far exceed the reduction observed in self-supervised settings. The presence of noise naturally results in larger error rates; further investigations are needed to demonstrate the transferability of noise robustness in a similar manner.

Acknowledgement

This work was supported by grant no. 31T139 at United Arab Emirates University and partially funded under UAEU-ZU Joint Research Grant G00003715 (Fund No.: 12T034) through Emirates Center for Mobility Research.

References

Ossama Abdel-Hamid, Li Deng, Dong Yu, and Hui Jiang. 2013. Deep segmental neural networks for speech recognition.

Hanan Aldarmaki, Asad Ullah, Sreepratha Ram, and Nazar Zaki. 2022. Unsupervised automatic speech recognition: A review. *Speech Communication*.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *arXiv preprint arXiv:2105.11084*.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020a. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

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

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

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *Proc. Interspeech 2020*, pages 5036–5040.

Yanzhang He and Eric Fosler-Lussier. 2015. Segmental conditional random fields with deep neural networks as acoustic models for first-pass word recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

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

Christiaan Jacobs, Yevgen Matusevych, and Herman Kamper. 2021. Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 919–926. IEEE.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. 2020. Multilingual acoustic word embedding models for processing zero-resource languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6414–6418. IEEE.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1182–1192.

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

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017.

Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proc. Interspeech 2017*, pages 498–502.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Puyuan Peng, Herman Kamper, and Karen Livescu. 2020. A correspondence variational autoencoder for unsupervised acoustic word embeddings. *Advances in neural information processing systems*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

Lisa van Staden and Herman Kamper. 2021. A comparison of self-supervised speech representations as input features for unsupervised acoustic word embeddings. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 927–934. IEEE.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

A Appendix

A.1 Dataset Details

A.2 Model Architecture & Hyper-Parameters

The architecture described in section 4.1 was modeled after other acoustic word embedding models (Chung et al., 2016; Chung and Glass, 2018; Holzenberger et al., 2018) with slight variations in details. We found that this particular configuration worked best across different acoustic features, whereas other choices gave mixed results. For example, using GRUs instead of LSTMs worked well with pre-trained features but was worse for MFCCs. The decoding process described in Holzenberger et al. (2018), where positional encodings are used instead of previous outputs also resulted in inferior performance. We also found that using teacher forcing instead of the model's previous output as input to the decoder hurt the performance. Finally, using two layers was crucial to get results in line with DTW perforamnce for most self-supervised models. The only exception is the self-supervised model with XLSR features which resulted in unstable training with 2 layers. We found it to work much better with a single layer network and slightly larger embedding size. Generally, larger embeddings sizes improved performance to some extent, but the improvements were smaller beyond the values that we have chosen; furthermore, using smaller sizes is more advantageous in terms of computational efficiency. We did not perform any hyper-parameter tuning for the target languages since we are working within the premise of low-resource settings where validation data may not be available.

Dataset	test	dev
English	52,576	54,402
French	90,958	83,560
German	121,713	122,903
Spanish	88,417	87,417
Arabic	62,745	57,532

Table 3: Total number of words in each dataset

Table 4 shows the number of parameters for each model. Since the decoder is only used for training and can be discarded after that, we only show the number of encoder parameters.

A.3 Training Details

The supervised models were trained with NLL loss, and the training targets are sequences of phonemes obtained using the Phonemizer package [6] (Bernard and Titeux, 2021). This choice seemed more sensible at first, but we found that using sequences of characters instead of phonemes worked equally well.

The model was implemented using PyTorch and trained on NVIDIA K80 GPU as provided in AWS p2.xlarge instances. For optimization, we found

[6] https://github.com/bootphon/phonemizer

Model	input	hidden	no.of parameters
Self-Supervised			
MFCC	39	100	354,400
Wav2Vec	768	100	937,600
MCPC	256	100	528,000
XLSR-53	1024	250	1,411,200
Supervised			
MFCC	39	100	354,400
Wav2Vec	768	100	937,600
MCPC	256	100	528,000
XLSR-53	1024	100	1,142,400

Table 4: Input size, hidden layer size, and total number of encoder parameters for each model.

that adam optimizer worked for all features except MFCCs, for which SGD with cyclical or step learning rate schedule was more stable.

Table 3 shows the number of words in each dataset. The word alignments were obtained via force alignment using The Montreal Forced Aligner[7] (McAuliffe et al., 2017) for English, German, French, and Spanish. The Montreal aligner uses an ASR engine, and since these datasets are relatively clean, the alignments are generally accurate. For Arabic, the best option was the aeneas toolkit[8], which relies on a TTS engine to align the synthesized words with the actual audio segments. We used Amazon Polly TTS for higher quality, but overall the alignments were not as accurate as the other datasets, which we believe is due to the low quality of the recordings, presence of noise, and high variability in accents. The low clustering accuracy could be partially attributed to the inaccurate labeling of the segments as a result of this. For ABX evaluation on the Arabic set, we manually filtered the segments that had somewhat accurate boundaries; the chosen pairs still contained high level of noise conditions, such as background music and interfering speech.

[7] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[8] www.readbeyond.it/aeneas