Unsupervised acoustic unit discovery by leveraging a language-independent subword discriminative feature representation

Feng, Siyuan; Zelasko, Piotr; Moro-Velázquez, Laureano; Scharenborg, Odette

DOI
10.21437/Interspeech.2021-1664

Publication date
2021

Document Version
Final published version

Published in
22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021

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

Unsupervised Acoustic Unit Discovery by Leveraging a Language-Independent Subword Discriminative Feature Representation

Siyuan Feng[1], *Piotr Żelasko*[2,3], *Laureano Moro-Velázquez*[2] *and Odette Scharenborg*[1]

[1]Multimedia Computing Group, Delft University of Technology, The Netherlands
[2]Center for Language and Speech Processing, Johns Hopkins University, USA
[3]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{S.Feng, O.E.Scharenborg}@tudelft.nl, {petezor,laureano}@jhu.edu

Abstract

This paper tackles automatically discovering phone-like acoustic units (AUD) from unlabeled speech data. Past studies usually proposed single-step approaches. We propose a two-stage approach: the first stage learns a subword-discriminative feature representation, and the second stage applies clustering to the learned representation and obtains phone-like clusters as the discovered acoustic units. In the first stage, a recently proposed method in the task of unsupervised subword modeling is improved by replacing a monolingual out-of-domain (OOD) ASR system with a multilingual one to create a subword-discriminative representation that is more language-independent. In the second stage, segment-level k-means is adopted, and two methods to represent the variable-length speech segments as fixed-dimension feature vectors are compared. Experiments on a very low-resource Mboshi language corpus show that our approach outperforms state-of-the-art AUD in both normalized mutual information (NMI) and F-score. The multilingual ASR improved upon the monolingual ASR in providing OOD phone labels and in estimating the phone boundaries. A comparison of our systems with and without knowing the ground-truth phone boundaries showed a 16% NMI performance gap, suggesting that the current approach can significantly benefit from improved phone boundary estimation.
Index Terms: Acoustic unit discovery, unsupervised subword modeling, zero-resource

1. Introduction

There are around $7,000$ spoken languages in the world [1], most of which lack transcribed speech data [2]. Conventional supervised acoustic modeling strategies [3,4] therefore cannot be applied directly to build ASR systems for such low-resource languages. As a result, current high-performance ASR schemes are available only for a very small number of languages [5]. To facilitate ASR for low-resource languages, unsupervised acoustic modeling has been gaining research interest recently [6–8]. Unsupervised acoustic modeling aims to discover basic speech units that represent all the sounds in a target language by making a *zero-resource* assumption [9], i.e., for a target language, only speech recordings are available while transcriptions and phoneme inventory (or its size) information are unknown.

There are two mainstream research strands in unsupervised acoustic modeling. The first strand, *acoustic unit discovery (AUD)* [6,10], formulates the problem as discovering a finite set of phone-like acoustic units [6,7,11]. The second strand, *unsupervised subword modeling (USM)* [9,12], formulates the problem as learning a frame-level feature representation that can distinguish subword units (phonemes) and is robust to speaker variation [8, 13, 14]. Studies on the USM task were mostly driven by the ZeroSpeech Challenges [9, 12, 15]. In essence, the USM task can be considered as learning an intermediate representation towards achieving the goal of AUD [16].

This study addresses the AUD task. Two main types of approaches to the AUD task were investigated in the past. The first type adopts self-supervised learning algorithms and uses a quantization layer to obtain a finite set of discovered acoustic units [13, 17, 18]. The second type adopts Bayesian non-parametric versions of the hidden Markov model (HMM) [6, 11, 19]. The combination of self-supervised learning and Bayesian approaches was also studied [20, 21]. All the studies mentioned above proposed single-step approaches. In contrast, the present study proposes a two-stage learning framework: the first stage learns a frame-level subword-discriminative feature representation (i.e., the USM task); the second stage applies clustering techniques to the learned frame representation to obtain a set of clusters as the discovered acoustic units. Subword-discriminative feature representations can provide a better separation between sounds than spectral features: In a subword-discriminative representation, two examples of the same phoneme are closer while those of different phonemes are further apart than in an MFCC representation. This is a highly desired property in clustering-based acoustic unit discovery [7, 22], which motivates us to propose a two-stage learning framework.

Specifically, in the first stage of the framework proposed in this study, we leveraged and improved a USM approach from a previous study [23]. This approach trains an autoregressive predictive coding (APC) model [24] followed by a cross-lingual DNN model to extract bottleneck features (BNFs) as the subword-discriminative representation. Previous results [23, 25] employing a *single* out-of-domain (OOD) language's resources to generate OOD phone labels for cross-lingual DNN training, provided state-of-the-art performance in USM tasks. Here, we aim to improve this approach further and, for the first time, use it for a different task: AUD. To leverage an OOD ASR system to generate OOD phone labels, We propose to use *multiple* OOD languages' resources to build a more language-independent OOD ASR system than in our previous work [23]. Because different languages have different phoneme inventories, we hypothesize that OOD phone labels that capture a more extensive set of sounds will be more useful for acoustic modeling of a target language. We will compare the use of multiple versus a single OOD language resources in this paper. In the second stage of our framework, the k-means algorithm is adopted for speech segment clustering. To that end, first, phone segment boundaries are estimated using an OOD ASR via de-

Code: https://github.com/syfengcuhk/mboshi.

http://dx.doi.org/10.21437/Interspeech.2021-1664

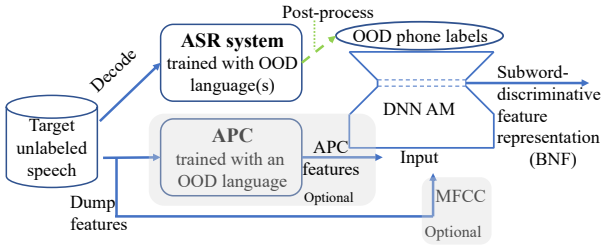

Figure 1: *The first stage in the proposed approach. "OOD" denotes out-of-domain, i.e. non-target language(s). The input to the DNN AM is either APC features or MFCCs but not both.*

coding [26]. The resulting variable-length segments need to be represented as fixed-dimension vectors, for which we compare two often adopted approaches [22, 26, 27]: an average-based method [7] and a downsampling method [28]. We measure the sensitivity of our AUD approach's performance to the number of discovered acoustic units, because the phone inventory size of a target language is usually unknown. The experiments are carried out on a very low-resource language, Mboshi [29] (4.5 hours of unlabeled speech).

2. Proposed two-stage approach

2.1. Stage 1: subword-discriminative feature learning

The goal of this stage is to learn a frame-level subword-discriminative feature representation. The general framework of the first stage of our proposed approach, based on [23], consists of an 1) APC model, which creates the APC features that are used as input to a 2) DNN acoustic model (AM) together with the frame-level phone labels obtained from 3) an OOD (non-target language) ASR system (see Figure 1). After training the DNN AM, BNFs are extracted from the bottleneck layer of the DNN as the desired subword-discriminative representation.

APC is a self-supervised learning model without the need of transcriptions for training. It is trained to predict a future speech frame n steps ahead (named *prediction step*) based on the current and past frames of an utterance [24]. APC is incorporated in [23] to extract *APC features* as input to the DNN AM (see Figure 1), owing to its ability to make phonetic and speaker information in speech more separable than MFCC. In this study we tackle an extremely low-resource scenario (4.5 h), however our previous work [23] suggests that a larger training dataset (more than 50 h) is needed to obtain effective APC features, thus we opt for using an APC model trained with a well-resourced OOD language. Note that a recent study [30] found that self-supervised models trained on one language can be used to represent another language with certain success. In our experiments we compare (1) APC features extracted by an APC model which is trained with an OOD language; and (2) MFCC features (hence bypassing the APC model); as the input to the DNN AM (see Figure 1).

The DNN AM in Figure 1 is trained with target language acoustic data. Frame-level phone labels required for training the DNN AM are obtained using an OOD (non-target) ASR system [23]: a target speech utterance is decoded by the OOD ASR system so that every frame is assigned a phone label generated by the OOD ASR. By this means, an OOD language's phonetic knowledge is exploited for the target language acoustic modeling. BNFs are then extracted from the bottleneck layer of the trained DNN as the desired subword-discriminative representation.

The language-independent OOD ASR system leverages multiple phonetically diverse languages' resources. We use International Phone Alphabet (IPA) symbols [31] to represent the phonemes of the different OOD languages, thus creating a phoneme inventory of the multilingual ASR system that is more language-independent [32] than that of a monolingual OOD. This also enables acoustic information sharing of the same or similar sounds from multiple languages during ASR training. A multilingual ASR system captures a wider phonetic space and has more different phone labels than a monolingual ASR system, thus is expected to provide more refined OOD phone labels for the target speech than a monolingual ASR system.

Another modification to the approach in [23] is adding a post-processing step to the OOD ASR based phone labels (see Figure 1). Essentially, the post-processing aims to refine the phone labels from an OOD ASR via re-aligning: First, we train an HMM with the OOD phone labels and Mboshi acoustic data; second, we generate the HMM phone alignments as the desired frame-level label supervision (rather than the output of an OOD ASR) to train the DNN AM. We experimentally found that the post-processing step consistently improves the AUD performance. Presumably, it refines the OOD phone labels by leveraging contextual information in the Mboshi acoustic data.

2.2. Stage 2: speech segment representation and clustering

This stage applies k-means clustering to the subword-discriminative feature representation learned by the first stage to obtain a finite set of clusters, each of which resembles a phone-like acoustic unit. The discovered units are the outcome of the proposed two-stage approach.

Speech clustering can be realized at the segment level [33] or at the frame level [8]. For segment-level clustering, we need the phone segment boundaries, and the segments need to be represented as fixed-dimension vectors. In order to obtain the segment boundaries, we rely on the OOD ASR system (see Section 2.1): after decoding the target speech data, phone boundary information is obtained by finding discontinuities of the frame-level OOD phone labels. This phone boundary estimation method is similar to [26], except that here we are using a multilingual and IPA symbol-based OOD ASR system.

This study compares two methods to obtain the fixed-dimensional segment representation. The first is a **downsampling** method [28] as suggested by [22, 27]: a variable-length speech segment is cut into a fixed number (s) of consecutive sub-segments, and the averages over d-dimensional frame-level feature vectors within each sub-segment are concatenated to form a feature vector of dimension $s \times d$ for each segment. Note that when $s = 1$, the method is equivalent to an **average**-based method [7] which takes the average of the frame-level features over all the frames in a segment. The downsampling method with a large s captures abundant temporal information which is not captured by the average method, however a large s leads to a high dimension of the segment-level feature vector which might adversely affect k-means.

Segment-level clustering with the two methods mentioned above are compared with a frame-level clustering system as a baseline, which applies k-means (same as in the proposed systems) to a frame-level feature representation. While circumventing the need for segment boundaries and the need for fixed-length segment representations, frame-level clustering tends to produce over-fragmented discovered units [34, 35]. Finally, we report an "upperbound" segment-level system by assuming the availability of golden phone boundary information while keep-

ing the other settings unchanged. This allows us to quantify the performance degradation attributed to imperfect phone boundary estimation.

3. Experimental setup

3.1. Evaluation metrics

We use two common metrics in the AUD task [10, 19, 20] to compare with past studies: normalized mutual information (**NMI**) and **F-score**. NMI measures the statistical dependency between discovered units (DUs) and ground-truth phone units (GUs), which is computed based on a frame-level confusion matrix of DU and GU (see [19] for details). An NMI value ranges between 0 and 100%, with a higher value indicating a higher consistency between DUs and GUs, hence is preferred. F-score is the harmonic mean of recall (R) and precision (P). It is used to measure the accuracy of the phone segmentation. A tolerance of ± 20 ms is set when computing F-score values. Higher F-score, R and P are preferred.

3.2. Databases

The AUD performance is evaluated on a corpus containing $5,130$ sentences spoken by three speakers of the Mboshi language [29], for a total amount of 4.5 hours. Automatically generated Mboshi phone alignments are available, but are not used during system development. The DNN AM in our system is trained and evaluated on the entire Mboshi dataset without training-test partition, as we are tackling an unsupervised learning problem. This is also consistent with the studies [11, 19].

Speech from 13 phonetically diverse languages [32] is used to train the OOD multi-/monolingual ASR systems. 5 languages are from GlobalPhone [36]: Czech (24 h), French (23 h), Spanish (12 h), Mandarin (15 h) and Thai (23 h). The other 8 languages are from IARPA Babel: Cantonese (127 h), Bengali (55 h), Vietnamese (78 h), Lao (59 h), Zulu (54 h), Amharic (39 h), Javanese (41 h) and Georgian (45 h).

3.3. Implementation of stage 1

The APC model included in the first stage of the proposed model is taken from our previous study [23]: it has 5 LSTM layers of dimension 100 with residual connections. The prediction step is 5. The model was trained with the Libri-light (English) *unlab-600 (hour)* set [37]. APC features are extracted from the top layer of the APC model.

We developed 2 multilingual systems and 5 monolingual systems, differing only in the training languages: **Multi-5** denotes the multilingual system trained with the 5 GlobalPhone languages; **Multi-13** denotes the multilingual system trained with all 13 GlobalPhone+Babel languages; **Mono-CZ, Mono-FR, Mono-SP, Mono-MA, Mono-TH** are five monolingual systems, each trained with one GlobalPhone language, i.e., Czech, French, Spanish, Mandarin, and Thai, respectively. We do not report any monolingual system trained on each of the Babel languages because of its inferior AUD performance – possibly explained by the large recording condition mismatch between the Babel languages and Mboshi. IPA symbols are used to represent the basic acoustic units, and the mapping from orthographic transcriptions to IPA symbol sequences is obtained by LanguageNet G2P models [38].

The multilingual and monolingual OOD ASR systems are trained using Kaldi [39], adopting a hybrid architecture [3], following implementation in [5]. The AM adopts a factorized time-delay neural network (TDNNF) consisting of 12 layers with a hidden dimension of 1024 and Resnet-style skip connections, trained with the LF-MMI criterion [40] for 4 epochs, with a starting learning rate (LR) of 10^{-3}. The input features consist of 43-dimension high-resolution MFCC+pitch features and 100-dimension i-vectors. The language model (LM) is a uni-gram phonotactic LM instead of an RNNLM, as we intend the OOD ASR phone labeling process to be minimally affected by the OOD language phonotactics. The LM is trained with the training data transcripts using SRILM [41].

The DNN AM for Mboshi is trained using either APC features or MFCC features of the Mboshi data. For each of the 7 multi/monolingual OOD ASR systems, the generated and post-processed (see Section 2.1) OOD phone labels are used to train one DNN AM with MFCC as input features, resulting in 7 DNN AMs. For the sake of simplicity, to test the effectiveness of the APC features, only for the system employing **Multi-13** (the best-performing OOD ASR), an additional DNN AM is trained with APC features instead of MFCCs. The Mboshi DNN AM adopts a TDNNF structure similar to that of the OOD ASR AM, except: a 40-dimension bottleneck layer is placed below the top TDNNF layer; i-vector input is not included as we found it deteriorated the performance in our preliminary results (not included in this paper); the model is trained for 20 epochs with a smaller LR of $2.5 \cdot 10^{-4}$ to stabilize the training procedure due to only 4.5 hours of training material. After training the DNN AM, the BNF for the Mboshi representations are extracted from the bottleneck layer as the learned frame-level subword-discriminative representation, and are used as input to the second stage of the proposed system.

3.4. Implementation of stage 2

The k-means algorithm is implemented using [42]. Unless specified differently, the number of clusters is empirically set to 50. Segment-level clustering is done on all the learned subword-discriminative representations of the different DNN AMs (i.e., 1 for each OOD system). For segment-level clustering, the speech segment boundaries are estimated using the OOD ASR system in the first stage. The downsampling method with s in $\{2, 3, 4, 5\}$ and the average based method are compared for obtaining the fixed-dimension segment representation. The frame-level clustering baseline and the upperbound segment-level system are based on the feature representation learned using **Multi-13**. Since the optimal setting of the number of clusters is unknown, we tested a range between $30 - 70$.

4. Results and discussion

For each experiment, we repeat k-means clustering 5 times with different random initialization and report NMI and F-score in means $\pm$ standard deviation.

4.1. Evaluation of stage 1: Effect of frame-level subword-discriminative feature representations

We first evaluate the effect of using a multilingual vs. a monolingual OOD ASR system on the effectiveness of stage 1. Next, we investigate the effect of using APC features as input features to the DNN AM. A fixed setting of stage 2 is used in all these experiments, i.e., the segment-level k-means with the average-based method to obtain the segment-level feature representation. The performances of our systems and two state-of-the-art (SotA) systems from the literature [11, 19] are listed in Table 1. Several observations can be made from this table:

Table 1: *Comparison of adopting a multi-/monolingual OOD ASR system in stage 1 of our approach and SotA [11, 19].*

Input	system	NMI (%)	F-score (%)
MFCC	Mono-CZ	40.87 ± 0.14	63.01 ± 0.06
	Mono-FR	38.32 ± 0.17	$\mathbf{64.14 \pm 0.10}$
	Mono-SP	37.57 ± 0.51	58.87 ± 0.08
	Mono-MA	38.85 ± 0.24	61.45 ± 0.12
	Mono-TH	37.61 ± 0.08	61.79 ± 0.05
	Multi-5	41.93 ± 0.28	62.84 ± 0.03
	Multi-13	$\mathbf{43.00 \pm 0.12}$	62.89 ± 0.07
APC	Multi-13	42.15 ± 0.28	62.90 ± 0.15
N/A	Yusuf et al. [19]	41.07 ± 1.09	59.15 ± 1.51
	Ondel et al. [11]	38.38 ± 0.97	59.50 ± 0.78

Table 2: *Speech clustering strategies in stage 2 of the proposed approach. "Seg./Fra." denotes segment- and frame-level clustering. "AVG‡" indicates the upperbound system.*

Type	System	NMI (%)	F-score (%)	Recall (%)	Precision (%)
Seg.	AVG	$\mathbf{43.00 \pm 0.12}$	$\mathbf{62.89 \pm 0.07}$	73.47	54.97
	DS-2	$\mathbf{43.00 \pm 0.12}$	62.87 ± 0.07	74.22	54.54
	DS-3	42.73 ± 0.28	62.70 ± 0.15	74.12	54.32
	DS-4	42.49 ± 0.27	62.47 ± 0.10	73.74	54.19
	DS-5	42.44 ± 0.16	62.60 ± 0.07	74.07	54.20
	AVG‡	59.29 ± 1.17	97.73 ± 0.06	100.00	100.00
Fra.	Baseline	41.82 ± 0.20	43.59 ± 0.35	90.38	28.72

Table 3: *NMI (row 2) and F-score (row 3) performances w.r.t different numbers of clusters (row 1).*

30	40	50	60	70
41.35 ± 0.21	42.50 ± 0.34	43.00 ± 0.12	$\mathbf{43.22 \pm 0.52}$	43.20 ± 0.53
62.81 ± 0.05	$\mathbf{62.89 \pm 0.04}$	$\mathbf{62.89 \pm 0.07}$	62.77 ± 0.14	62.76 ± 0.10

(1) The multilingual OOD ASR systems (Multi-13/Multi-5) outperform the monolingual systems on the NMI measure. Apparently, using a language-independent OOD ASR system to provide the OOD phone labels is better than using a language-dependent system for target language DNN AM training, to learn a better frame-level subword-discriminative feature representation. Looking at the F-score, the Multi-5 and Multi-13 systems perform better than the average over the 5 monolingual systems (61.85%), nevertheless Mono-FR achieves the best performance, followed by Mono-CZ. The results indicate that using a multilingual OOD ASR is more beneficial for improving the phonetic relevance of the discovered acoustic units (NMI) than for improving phone boundary estimation (F-score).

(2) Our best system (Multi-13 with MFCC input) outperforms state-of-the-art [11,19] on both NMI and F-score. Similar to our approach, [11, 19] also relied on OOD languages' transcribed data for model training. However, they used around 35 hours (from 7 languages), while our two best systems (in NMI) used more OOD speech data - Multi-13: 595 hours; Multi-5: 97 hours. Nevertheless, our Mono-CZ system performs on par with or better than [11, 19] in NMI and F-score respectively, while using only 24 hours of Czech data.

(3) Comparison of the two Multi-13 systems shows that APC features as input to the DNN AM in stage 1 does not affect the F-score and deteriorates the NMI performance compared to MFCCs. This result, seemingly in contrast to [23], can be explained by the following: in [23], the APC model was trained on the target language English, while here the English-trained APC model was used to capture the target language (Mboshi). Moreover, [23] showed the success of APC features for the USM task, here we used a different, AUD task.

4.2. Evaluation of stage 2: Clustering strategies

We investigated the effect of the two segment representation strategies in stage 2, i.e., the downsampling method with different s and the average-based method, and compared these to the frame-level baseline and the system with access to golden phone boundary information (upperbound: "AVG‡"). The Multi-13 OOD ASR system is used to generate the OOD phone labels in all experiments; APC is not adopted. Table 2 shows the results. "AVG" denotes the average-based method, "DS-2$\sim$5" denotes the downsampling method with $s = 2 \sim 5$. In addition to NMI and F-score, Table 2 reports the average *recall* and *precision* values for each system, in order to gain deeper insights into the differences between segment- and frame-level k-means.

It can be clearly seen that the systems adopting segment-level k-means outperform the frame-level baseline on NMI and F-score. The superiority of the segment-level systems is more prominent on the F-score (absolute 19.0%) than on NMI (absolute 1.2%). Particularly, the baseline model has a very low *precision*, indicating a large proportion of false boundaries that are hypothesized. This implies a frame-level system tends to over-segment target speech, which is in line with [34, 35].

Table 2 shows that the downsampling method does not have an advantage over a simpler, average based method, and a larger s leads to a slight NMI degradation. While other studies showed a good performance for the downsampling method [22, 27], we show that in this low-resource Mboshi database, using the k-means algorithm, the average based method is comparable, if not better than the downsampling method.

Finally, Table 2 shows that the upperbound system (AVG‡) outperformed our best system (AVG) by 16.3% absolute NMI. The NMI gap is attributed exclusively to the imperfect phone boundary estimation by the OOD ASR system. It is expected that by improving the phone boundary estimation, or by adopting an interactive approach to refining segmentation and target language acoustic modeling [27], the frame-level subword-discriminative representation learned in the first stage could be based on to achieve an NMI that approaches the upperbound.

The effect of the number of clusters on the AUD task was investigated using the average based method. Table 3 summarizes the results: the best NMI is obtained with a number of clusters between 60 and 70. The F-score performance is less sensitive to the number of clusters than NMI. Overall, a number of clusters between 50 and 70 shows the best performance.

5. Conclusions and future work

This paper proposes a two-stage approach for the unsupervised AUD task. Our best model, which employs 13 OOD language resources in stage 1 to provide phone labels for target language AM training and uses an average-based method segment clustering in stage 2, outperforms state-of-the-art performance on a very low-resource Mboshi database. The results showed that the multilingual OOD ASR systems outperformed a monolingual one in providing the frame labels for target language acoustic modeling and in phone boundary estimation, with the former being more prominent. Comparison with a golden standard showed that a 16.0% NMI performance gap could be attributed to imperfect phone boundary information. Furthermore, in stage 1, APC features were compared to MFCCs as input to the DNN AM training module and were less effective in the AUD task. In stage 2, the best performance was achieved using a number of clusters between 50 and 70.

6. Acknowledgements

We thank Lucas Ondel for valuable discussion on the evaluation software.

7. References

[1] P. K. Austin and J. Sallabank, *The Cambridge handbook of endangered languages*. Cambridge University Press, 2011.

[2] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon *et al.*, "Breaking the unwritten language barrier: The bulb project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[3] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE TASLP*, vol. 20, no. 1, pp. 30–42, 2011.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[5] S. Feng, P. Zelasko, L. Moro-Velázquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "How phonotactics affect multilingual and zero-shot ASR performance," *To appear in ICASSP*, 2021.

[6] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *ACL*, 2012, pp. 40–49.

[7] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Acoustic segment modeling with spectral clustering methods," *IEEE/ACM Trans. ASLP*, vol. 23, no. 2, pp. 264–277, 2015.

[8] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *INTERSPEECH*, 2015, pp. 3189–3193.

[9] E. Dunbar, X.-N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier *et al.*, "The zero resource speech challenge 2017," in *ASRU*, 2017, pp. 323–330.

[10] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for acoustic unit discovery," *SLTU*, vol. 81, pp. 80–86, 2016.

[11] L. Ondel, H. K. Vydana, L. Burget, and J. Cernocký, "Bayesian subspace hidden Markov model for acoustic unit discovery," in *INTERSPEECH*, 2019, pp. 261–265.

[12] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao *et al.*, "The zero resource speech challenge 2019: TTS without T," in *INTERSPEECH*, 2019, pp. 1088–1092.

[13] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Advances in NIPS*, 2017, pp. 6306–6315.

[14] M. Heck, S. Sakti, and S. Nakamura, "Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017," in *ASRU*, 2017, pp. 740–746.

[15] E. Dunbar, J. Karadayi, M. Bernard, X.-N. Cao, R. Algayres, L. Ondel *et al.*, "The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units," in *INTERSPEECH*, 2020, pp. 4831–4835.

[16] S. Feng, T. Lee, and Z. Peng, "Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling," in *INTERSPEECH*, 2019, pp. 1093–1097.

[17] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[18] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge," in *INTERSPEECH*, 2020, pp. 4836–4840.

[19] B. Yusuf, L. Ondel, L. Burget, J. Cernocký, and M. Saraclar, "A hierarchical subspace model for language-attuned acoustic unit discovery," *CoRR*, vol. abs/2011.03115, 2020.

[20] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj, "Hidden markov model variational autoencoder for acoustic unit discovery," in *INTERSPEECH*, 2017, pp. 488–492.

[21] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg *et al.*, "Bayesian models for unit discovery on a very low resource language," in *ICASSP*, 2018, pp. 5939–5943.

[22] S. Bhati, S. Nayak, K. S. R. Murty, and N. Dehak, "Unsupervised Acoustic Segmentation and Clustering Using Siamese Network Embeddings," in *INTERSPEECH*, 2019, pp. 2668–2672.

[23] S. Feng and O. Scharenborg, "Unsupervised Subword Modeling Using Autoregressive Pretraining and Cross-Lingual Phone-Aware Modeling," in *INTERSPEECH*, 2020, pp. 2732–2736.

[24] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *INTERSPEECH*, 2019, pp. 146–150.

[25] S. Feng and O. Scharenborg, "The effectiveness of unsupervised subword modeling with autoregressive and cross-lingual phone-aware networks," *Under review by IEEE OJ-SP. Manuscript available at ArXiv.*, vol. abs/2012.09544, 2020.

[26] S. Feng, T. Lee, and H. Wang, "Exploiting language-mismatched phoneme recognizers for unsupervised acoustic modeling," in *ISCSLP*, 2016, pp. 1–5.

[27] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," in *ASRU*, 2017, pp. 719–726.

[28] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *ASRU*, 2013, pp. 410–415.

[29] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt *et al.*, "A very low resource language speech corpus for computational language documentation experiments," in *LREC*, 2018.

[30] M. Rivière, A. Joulin, P. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020, pp. 7414–7418.

[31] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[32] P. Żelasko, L. Moro-Velázquez, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "That sounds familiar: an analysis of phonetic representations transfer across languages," in *INTERSPEECH*, 2020.

[33] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *ICASSP*, 1988, pp. 501–504.

[34] B. Wu, S. Sakti, J. Zhang, and S. Nakamura, "Optimizing DPGMM clustering in zero-resource setting based on functional load," in *SLTU*, 2018, pp. 1–5.

[35] S. Feng and T. Lee, "Exploiting cross-lingual speaker and phonetic diversity for unsupervised subword modeling," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2000–2011, 2019.

[36] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university," in *INTERSPEECH*, 2002.

[37] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré *et al.*, "Libri-light: A benchmark for ASR with limited or no supervision," in *ICASSP*, 2020, pp. 7669–7673.

[38] M. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G.-A. Levow, and K. Kirchhoff, "Grapheme-to-phoneme transduction for cross-language asr," in *ICSLSP*, 2020, pp. 3–19.

[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[40] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH*, 2016, pp. 2751–2755.

[41] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *ICSLP*, 2002, pp. 901–904.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.