Vector Quantized Temporally-Aware Correspondence Sparse Autoencoders for Zero-resource Acoustic Unit Discovery

Batuhan Gundogdu[1,2], *Bolaji Yusuf*[1], *Mansur Yesilbursa*[1], *Murat Saraclar*[1]

[1]Bogazici University, Turkey
[2]National Defense University Naval Academy, Turkey

{batuhan.gundogdu, bolaji.yusuf, mansur.yesilbursa, murat.saraclar}@boun.edu.tr

Abstract

A recent task posed by the Zerospeech challenge is the unsupervised learning of the basic acoustic units that exist in an unknown language. Previously, we introduced recurrent sparse autoencoders fine-tuned with corresponding speech segments obtained by unsupervised term discovery. There, the clustering was obtained on the intermediate layer where the nodes represent the acoustic unit assignments. In this paper, we extend this system by incorporating vector quantization and an adaptation of the winner-take-all networks. This way, symbol continuity could be enforced by excitatory and inhibitory weights along the temporal axis. Furthermore, in this work, we utilized the speaker information in a speaker adversarial training on the encoder. The ABX discriminability and the low bitrate results of our proposed approach on the Zerospeech 2020 challenge demonstrate the effect of the enhanced continuity of the encoding brought by the temporal-awareness and sparsity techniques proposed in this work.

Index Terms: sparse recurrent autoencoder, correspondence autoencoder, winner-take-all nets, speaker adversarial training, vector quantization

1. Introduction

Existing speech technology heavily relies on annotated corpora. However, such corpora don't exist for many languages in the world. To address this problem, zero resource speech processing focuses on unsupervised methods to create linguistic and acoustic models for the low-resource languages [1]. Such systems are aimed to be trained on small amount of unlabeled data and independent of the language so that they are adaptable to many others.

Zerospeech Challenge provides a platform to test various zero-resource systems on the same task, thus creating valuable literature on the field. To continue this mission, Zerospeech Challenge 2019 task tackles the problem of developing a speech synthesis system without any written text data [2]. The task is specifically interesting because we know that infants are able to produce speech without being exposed to any annotated data. The problem can be broken into two seemingly independent tasks: discovering speaker independent discrete acoustic units and producing speech in the voice of target speakers by using acoustic units. In this paper, we focused on the acoustic unit discovery task of the challenge.

Several unsupervised methods have been proposed to discover acoustic units in the recent past. One prevalent approach to the problem is to use Dirichlet process Gaussian mixture model - Hidden Markov Model (DPGMM-HMM) [3, 4], or DPGMM [5, 6]. Another approach is to utilize autoencoders [7, 8] or Siamese-style neural networks trained with unsupervised spoken term discovery (STD) labels [9, 10].

Following the literature, most of the methods proposed for acoustic unit discovery in Zerospeech 2019 Challenge were based on the Bayesian non-parametric approach with DPGMM [11] or HMM-GMM [12] and neural network models with autoencoder architecture [13, 14, 15]. DPGMM structure is used to generate pseudo-labels for supervised deep neural network (DNN) training [11] whereas HMM-GMM framework is employed to discover acoustic unit representations from clusters of CVC segments in another work [12]. Autoencoder models are trained to obtain acoustic unit representations in the latent layer [13, 14, 15] or to reconstruct speaker-independent MFCC features before unit discovery [11]. Since the challenge requires discrete representation of the acoustic units, vector-quantized variational autoencoders (VQ-VAE) were utilized due to their ability to produce discrete latent variables [14, 15].

The methodology proposed in this paper is an extension of the correspondence recurrent sparse autoencoder (CoRSA) architecture which was introduced in [13] and has recently been applied to a similar task of discovering visual units in sign language videos [16]. In CoRSA, corresponding pairs of acoustically similar sequences are used to fine-tune a recurrent sparse autoencoder, which is similar to the methodology of [17]. The acoustic units to be discovered are obtained on the intermediate softmax layer. In this paper, we use an adaptation of the winner-take-all (WTA) networks that are more extensively used in computational models of the brain [18]. WTA networks are an example of competitive learning and generally used in recurrent form, where only one (particularly the strongest) of the nodes keep activating and others zero out. This feature makes them useful for distributed decision-making [19] applications. We adapt the WTA methodology to include a temporal aspect in this paper, and use it after the softmax layer of CoRSA. We apply vector quantization (VQ) on the output of one forward run of WTA, just to omit its recurrent application. In addition to WTA and VQ, which did not exist in [13], in this paper, we also incorporate speaker adversarial training (SAT) to the CoRSA model. The enforced continuity and unit consistency brought by WTA and SAT resulted in a significantly reduced bitrates. The proposed system obtained the lowest bitrate among the Zerospeech 2020 groups, while maintaining an ABX discriminability score outperforming the baseline on both languages.

2. System Description

The system follows the autoencoder-based approach that has been commonly adopted for the task of acoustic unit discovery. The proposed method is an enhancement of CoRSA which was previously proposed for the Zerospeech 2019 challenge. We'll briefly introduce the basic components of and the idea

behind CoRSA, then present the extensions to it, brought by the Zerospeech 2020 work in the following subsections.

2.1. Correspondence Recurrent Sparse Autoencoder

The proposed system is an autoencoder in which the underlying acoustic units that constitute the training speech data is expected to be obtained at the intermediate layer. The probability distribution over the set of acoustic units is obtained at the intermediate layer of the autoencoder as the output of a fully connected layer with softmax. The sparsity of this layer is ensured by maximizing the $L2$-norm of the layer while minimizing the reconstruction loss. This method is referred to as the recurrent sparse autoencoder (RSA), in [13] and in this paper. The consistency of the activations of the appropriate units on this layer is achieved by using pairs of acoustically similar sequences, obtained by unsupervised term discovery, in the reconstruction. These pairs of similar sequences, referred to as the corresponding pairs, are aligned with dynamic time warping (DTW). They are used as input-output pairs iteratively to the autoencoder. This training method is referred to as Correspondence RSA (CoRSA).

The encoder takes the speech frames $\mathbf{x} \in \mathcal{R}^{D\times T}$ as input, and feeds the clustering layer with the embedding $\mathbf{h} \in \mathcal{R}^{H\times T}$. The encoder can be in any form, i.e. fully connected, recurrent or convolutional. We used gated recurrent units for the encoder and the decoder. The sparse posteriorgram representation $\mathbf{p} \in [0,1]^{K\times T}$ is obtained at the output of the clustering layer, where K is the number of units. The reconstructed output $\hat{\mathbf{x}}$ is then obtained via a decoder. In CoRSA training, the network is fine-tuned with a lower learning rate, using corresponding pairs $\mathbf{x}$ and $\tilde{\mathbf{x}}$ as input and output pairs. The flowchart for the CoRSA model, and the proposed model is given in Figure 1. In this work, we extend the methodology of CoRSA with a temporal adaptation of winner-take-all networks, vector quantization and speaker adversarial training.

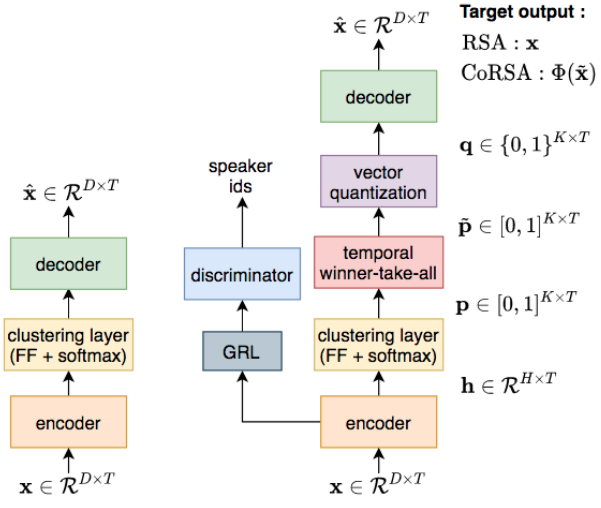

Figure 1: *CoRSA (left) and the proposed system (right)*

2.2. Winner-Take-All Network

The architecture and the training of CoRSA is very similar to other correspondence autoencoder-based [7, 8] and VQ-VAE-based [14, 15] systems. The main difference is that the intermediate layer activations are treated as the probability distribution of the acoustic units to be discovered. This is why we refer to this layer as the clustering layer. The sparsity loss of the RSA and CoRSA training reduces the entropy of this distribution, enabling consistent frame-unit assignments, at the risk of losing some of the K assignments completely. We take advantage of this aspect of the CoRSA method, by incorporating the winner-take-all (WTA) networks that are commonly used in computational models of the brain [18]. In WTA, unit activations are effected from inter-connected excitatory and inhibitory neural activations. More specifically, pseudo-unit activations on the clustering layer are fed to a WTA network with K input and output neurons. Each input neuron '*excites*' the corresponding output neuron with a positive weight, and '*inhibits*' other output neurons with negative weights. The idea is that eventually, the neuron with the maximum activation remains. In this paper, we extend the application of the WTA network to include temporal excitatory and inhibitory activations as well, in order to enforce continuity of the activations. Temporal adaptation of the WTA network is given in (1) and demonstrated in Figure 2.

$$\begin{aligned}\mathbf{p}_t &= \text{softmax}(\text{encoder}(\mathbf{x}_t))\\ \tilde{\mathbf{p}}_t^i &= \text{ReLU}(\alpha\mathbf{p}_t^i - \sum_{j\neq i}\beta\mathbf{p}_t^j + \gamma\mathbf{p}_{t-1}^i - \sum_{j\neq i}\psi\mathbf{p}_{t-1}^j)\\ \tilde{\mathbf{p}}_t &= \text{softmax}(\tilde{\mathbf{p}}_t)\end{aligned} \tag{1}$$

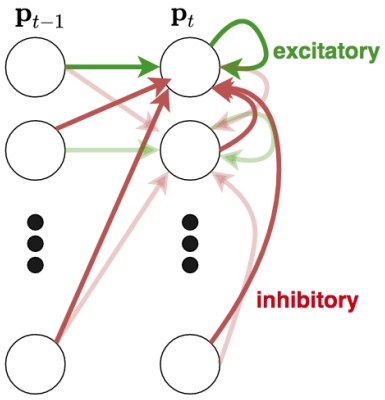

Figure 2: *Temporal WTA network.*

In (1), the hyper-parameters $\alpha, \gamma \geq 0$ are excitatory and $\beta, \psi \geq 0$ are inhibitory weights. It should be noted that, their relative power with respect to each other, rather than actual values, is important for determining the non-zero unit activation(s). Also, it can be seen from (1) that application of this layer repeatedly will result in a one-hot vector, with only the most probable unit activating, hence the name winner-take-all. We did not, however, implement a recurrent WTA network, instead, we quantized the output of the WTA network to the nearest one-hot representation:

$$\mathbf{q}_t^i = \begin{cases} 1, & \text{if } i = \arg\max_j \tilde{\mathbf{p}}_t^j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The decoder, which is taken as a stack of time-distributed dense and GRU layers, then uses the quantized unit estimations $\mathbf{q}$ as input to generate the reconstructed sequence $\hat{\mathbf{x}}$. While

training the encoder-decoder parameters, we use the sparsity loss of the layer right before the decoder in addition to the reconstruction loss, as proposed in [13]. For the CoRSA training, corresponding pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ are used as input-output pairs. We align the corresponding pairs with the DTW algorithm to have the same duration over DTW path (Φ). For RSA training, the ground truth is the input itself, i.e. $\Phi(\tilde{\mathbf{x}}_t) = \mathbf{x}_t$.

$$\begin{aligned} \hat{\mathbf{x}}_t &= \text{decoder}(\mathbf{q}_t) \\ \mathcal{L}_{CoRSA} &= \sum_{t \in 1 \cdots T} ||\Phi(\tilde{\mathbf{x}}_t) - \hat{\mathbf{x}}_t||^2 - \lambda ||\mathbf{q}_t||_2^2 \end{aligned} \tag{3}$$

It is obvious that the $L2$-norm of $\mathbf{q}$ is constant, but adding the negative of it to the total loss has the effect of enforcing the continuity of the reconstruction, and therefore a reduced bitrate. Since the input to the decoder is now discrete, and since speech is slow varying, to maintain that continuity, each frame before the decoder is enforced to have the same information as the previous frame, i.e. have the same unit activated.

2.3. Speaker Adversarial Training

As stated before, the Zerospeech training is meant to be unsupervised and the only information provided is the segmentation of utterances with respect to different speakers. In this work, we incorporate speaker adversarial training (SAT), as well, to CoRSA scheme in order to make use of the speaker information. The adversarial training approach adopted in this work is similar to [20] and [21]. The output of the encoder is directed to a discriminator that aims to identify the speaker information. As the discriminator block, we used a multi layer fully connected neural network followed by a softmax. The discriminator criterion ($\mathcal{L}_{disc}$) is the cross entropy between the speaker id labels and softmax(discriminator(encoder($\mathbf{x}$)). In order for the encoder to be speaker independent, the gradient is reversed between the encoder and the discriminator in the back propagation. This operation is depicted with a gradient reversal layer (GRL) on Figure 1. If we call the parameters of the encoder and the discriminator θ_{enc} and θ_{disc}, respectively, speaker adversarial the update rules can be given as follows:

$$\begin{aligned} \boldsymbol{\theta}_{disc} &\leftarrow \boldsymbol{\theta}_{disc} - \eta_1 \nabla_{\boldsymbol{\theta}_{disc}} \mathcal{L}_{disc} \\ \boldsymbol{\theta}_{enc} &\leftarrow \boldsymbol{\theta}_{enc} - \eta_1 \left(\nabla_{\boldsymbol{\theta}_{enc}} \mathcal{L}_{CoRSA} - \eta_2 \nabla_{\boldsymbol{\theta}_{enc}} \mathcal{L}_{disc} \right) \end{aligned} \tag{4}$$

In order to obtain the low bit-rate test embeddings, we employ a median filter along time temporal axis, after the WTA network, to avoid erroneous and unmeaningfully short state assignments on the encoding, similar to [13]. We then quantize the resulting embeddings to the nearest one-hot representation, as shown in (2).

$$\tilde{\mathbf{p}}_t^i = \text{median}(\tilde{\mathbf{p}}_{t-k:t+k}^i) \tag{5}$$

2.4. Implementation Details and Hyper-parameters

The models have been implemented with PyTorch and the source code has been made publicly available[1]. We used PLP features, normalized with the decaying exponential weigth of 0.9 as explained in [13]. For selecting the CoRSA pairs, we took the cutting threshold of UTD to be 0.92 and took the sequences that have a greater similarity score than this threshold. Both the encoder and decoder are a stack of GRU models and linear layers. The network is trained with Adam optimization [22]. We have observed that best results are obtained by training the network with RSA for some number of epochs, then running interleaving epochs of CoRSA and SAT. It should be noted that, changing the input-output pairs in CoRSA, i.e. generating $\mathbf{x}$ from $\Phi(\tilde{\mathbf{x}})$, as well as generating $\Phi(\tilde{\mathbf{x}})$ from $\mathbf{x}$ brings further improvements. The other hyperparameters of the system are given in Table 1.

[1] https://github.com/batuhan-gundogdu/corsa_wta/

Table 1: *Hyperparameters of the work described in this paper*

Parameter		Value
D	input dimensionality	16
H	Hidden layer size	128
K	Number of units	64
T	sequence length	250 (RSA), 80 (CoRSA)
N	batch size	1024 (RSA), 256 (CoRSA)
λ	sparsity weight	1.0
η_1	learning rate	0.0001 (RSA) 0.0005 (CoRSA)
η_2	adversarial cost	1.0
α	excitatory weight	$K-1$
β	inhibitory weight	1.0
γ	temporal excitatory weight	$K/2$
ψ	temporal inhibitory weight	0.0
k	Median filter order	3

3. Experiments

Zerospeech 2020 challenge had 3 tracks, which are the repetition of 2017 and 2019 challenges. The 2017 tracks were focused on unsupervised term discovery and unit discovery. The 2019 track was defined as the TTS without T task, hence it is also an unsupervised unit discovery task, but the compression of the frame representations are also measured, along with the synthesis using the discovered units. This work presents our work submitted to Zerospeech 2020, for the 2019 task. In this work, we focused on obtaining a minimum bitrate representation possible without harming the ABX discrimination performance.

The ABX discriminability of the produced embeddings is calculated via a DTW-based sequence comparison [23]. As a part of the challenge, the synthesized wave files are evaluated by the judges based on their intelligibility through mean opinion scores (MOS). For the synthesis we used the baseline system [24] and focused primarily on lowering bitrate and ABX error simultaneously. The experiments are conducted on the two languages provided in the challenge: English and the surprise language, later revealed to be Standard Indonesian [25, 26].

3.1. Experimental Results

The English dataset was provided as the system development set by the program. The ABX and bitrate results demonstrating the several components of the proposed method are given in Table 2. To begin with, we initially took the intuitive benchmarks as baselines for comparison. The "no compression" system demonstrates the ABX error when no compression and clustering is applied. It is naturally feasible to obtain better results than that without taking the bitrate into consideration, yet this is the goal of the 2017 track challenge. The main objective here is to reduce the ABX error along with the bitrate. Applying k-means clustering (with $K = 64$ to be comparable

with this work), and using the one-hot representation of the nearest centroid for the test set, reduces the bitrate more than 86%, yielding an ABX still better than the topline, which uses LVCSR system labels. It can be seen that both RSA and CoRSA bring improvements over the provided baseline system, which consists of an acoustic unit discovery system based on HMM with Dirichlet process priors [3]. In each row of Table 2, we present the contribution brought by each of the system components (WTA, VQ, SAT) introduced in this paper as an extension to CoRSA. The medFilt on the last system, which was also the system submitted to Zerospeech, denotes the temporal median filtering.

Table 2: *English set results of the system components*

Model	**ABX**	**Bitrate**
no compression (PLP features)	22.0	1216.3
k-means (on PLP features)	29.6	169.3
Zerospeech baseline [3]	35.6	72.0
RSA	32.1	58.4
CoRSA	31.5	58.1
CoRSA + WTA	30.3	44.2
CoRSA + WTA + VQ	29.2	44.7
CoRSA + WTA + VQ + SAT	29.4	38.5
CoRSA + WTA + VQ + SAT + medFilt	29.9	34.6
topline (LVCSR)	29.8	37.3

3.2. Comparison with other systems in Zerospeech

We submitted the system with the lowest bit-rate (CoRSA + WTA + VQ + SAT + medFilt) to Zerospeech 2020. Our systems' results along with other submissions to the challenge are demonstrated on the bitrate-ABX plot given in Figures 3 and 4, for English and the surprise language, respectively.

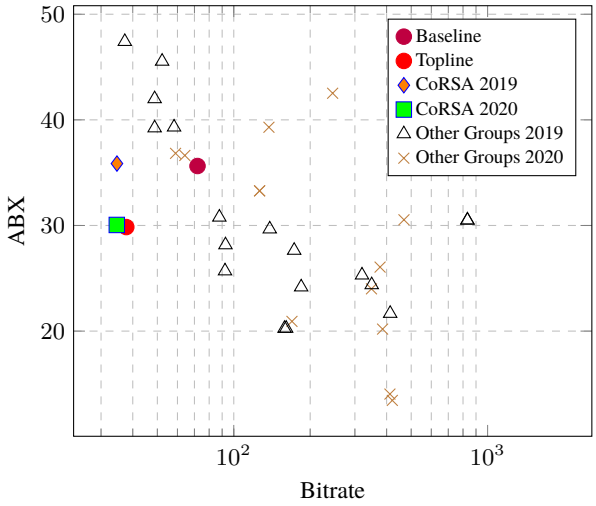

Figure 3: *ABX score vs bitrate in English*

The performance of the proposed system exhibits an obvious outlier behavior on both languages, as seen in Figures 3 and 4. More drastically, in English, the proposed system lands itself right next to the topline, outperforming the baseline, on both ABX and bitrate measures. There is also significant improvement in the ABX without an increase in bitrate compared to 2019 CoRSA system. This improvement is clearly brought by the WTA, VQ and SAT additions to the system.

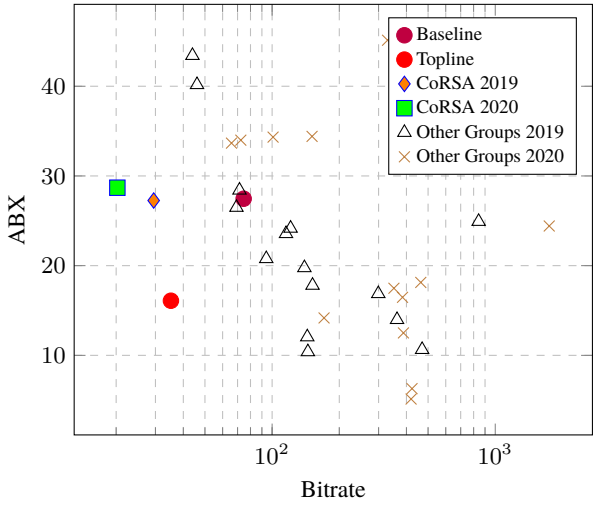

Figure 4: *ABX score vs bitrate in Surprise language*

Our system gives ABX scores close to its performance in English, with lower bitrate in surprise language. However, other systems, including topline, performed better on surprise language compared to English in terms of ABX and bitrate. Even though our system achieved the lowest bitrate, ABX score can be improved to obtain a system with higher discrimination ability.

4. Discussion and Conclusion

This paper describes our Zerospeech 2020 submission system. We primarily focused on the acoustic unit discovery task while minimizing the bitrate. The improvements made on to the CoRSA 2019 system include winner-take-all networks, vector quantization and speaker adversarial training. The experiments on the Zerospeech 2020 data show that the proposed improvements reduced the ABX error more than 5% absolute points without increasing the bitrate. We have shown that our system stands out from the general trend of ABX - bitrate relationship, achieving reasonably well ABX scores with very small bitrate in English.

5. Acknowledgements

This study was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project 116E076. Authors would like to thank Alican Gok, Oyku Deniz Kose and Korhan Polat for their invaluable help with the unsupervised term discovery and Nazif Can Tamer for their fruitful discussions.

6. References

[1] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, and E. Dupoux, "The Zero Resource Speech Challenge 2019: TTS without T," in *the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019): Crossroads of Speech and Language*, 2019.

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

[4] C. Y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[5] 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 *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] M. Heck, S. Sakti, and S. Nakamura, "Unsupervised linear discriminant analysis for supporting dpgmm clustering in the zero resource scenario," *Procedia Computer Science*, vol. 81, pp. 73–79, 2016.

[7] L. Badino, A. Mereta, and L. Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5818–5822.

[9] R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrum—deep siamese network pipeline for unsupervised acoustic modeling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4965–4969.

[11] S. Feng, T. Lee, and Z. Peng, "Combining adverserial training and disentangled speech representation for robust zero-resource subword modeling," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1093–1097.

[12] K. Pandia D S and H. A. Murthy, "Zero resource speech synthesis using transcripts derived from perceptual acoustic units," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1113–1117.

[13] B. Yusuf, A. Gok, B. Gundogdu, O. D. Kose, and M. Saraclar, "Temporally-aware acoustic unit discovery for zerospeech 2019 challenge," *Proc. Interspeech 2019*, pp. 1098–1102, 2019.

[14] A. Tjandra, B. Sisman, M. Zhang, S. Skati, H. Li, and S. Nakamura, "Vqvae unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1118–1122.

[15] R. Eloff, A. Nortje, B. v. Niekerk, A. Govender, L. Nortje, A. Pretorius, E. v. Biljon, E. v. d. Westhuizen, L. v. Staden, and H. Kamper, "Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1103–1107.

[16] R. D. Siyli, B. Gundogdu, M. Saraclar, and L. Akarun, "Unsupervised key hand shape discovery of sign language videos with correspondence sparse autoencoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8179–8183.

[17] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] Y. Fang, M. A. Cohen, and T. G. Kincaid, "Dynamics of a winner-take-all neural network," *Neural Networks*, vol. 9, no. 7, pp. 1141–1154, 1996.

[19] N. Lynch, C. Musco, and M. Parter, "Winner-take-all computation in spiking neural networks," *arXiv preprint arXiv:1904.12591*, 2019.

[20] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, "Speaker invariant feature extraction for zero-resource languages with adversarial learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2381–2385.

[21] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *Proc. Interspeech 2019*, pp. 3148–3152, 2019.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "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*, 2013, pp. 1–5.

[24] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 202–207. [Online]. Available: http://dx.doi.org/10.21437/SSW.2016-33

[25] S. Sakti, R. Maia, S. Sakai, T. Shimizu, and S. Nakamura, "Development of HMM-based indonesian speech synthesis," in *Proc. Oriental COCOSDA*, 2008, pp. 215–219.

[26] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of indonesian large vocabulary continuous speech recognition system within A-STAR project," in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, 2008.