INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

The Zero Resource Speech Challenge 2020: Discovering discrete subword and word units

Ewan Dunbar[1,2], *Julien Karadayi*[2], *Mathieu Bernard*[2], *Xuan-Nga Cao*[2], *Robin Algayres*[2], *Lucas Ondel*[3], *Laurent Besacier*[4], *Sakriani Sakti*[5,6], *Emmanuel Dupoux*[2,7]

[1]Université de Paris, LLF, CNRS, Paris, France
[2]Cognitive Machine Learning (ENS - CNRS - EHESS - INRIA - PSL Research University), France
[3]Department of Computer Graphics and Multimedia, Brno Univ. of Technology, Czech Republic
[4]Laboratoire d'Informatique de Grenoble, équipe GETALP (Université Grenoble Alpes), France
[5]Nara Institute of Science and Technology
[6]RIKEN Center for Advanced Intelligence Project, Japan
[7]Facebook A.I. Research, Paris, France

ewan.dunbar@utoronto.ca

Abstract

We present the Zero Resource Speech Challenge 2020, which aims at learning speech representations from raw audio signals without any labels. It combines the data sets and metrics from two previous benchmarks (2017 and 2019) and features two tasks which tap into two levels of speech representation. The first task is to discover low bit-rate subword representations that optimize the quality of speech synthesis; the second one is to discover word-like units from unsegmented raw speech. We present the results of the twenty submitted models and discuss the implications of the main findings for unsupervised speech learning.

Index Terms: zero resource speech technology, speech synthesis, acoustic unit discovery, spoken term discovery, unsupervised learning

1. Introduction

Current speech technology depends heavily on the availability of textual resources. On the other hand, humans learn the sounds and vocabulary of their first language long before they learn to read or write, discovering some kind of linguistic units or representations in their language (typically thought to be phoneme- or word-like), and the equivalent of an acoustic model, a language model, and a speech synthesizer. That humans succeed without textual resources suggests that there may be another approach. Developing technology to learn useful speech representations in an unsupervised way would be useful for the thousands of so-called low-resource languages, which lack the textual resources and/or expertise required to build traditional speech processing systems.

The Zero Resource Speech Challenge series[1] [1, 2, 3] aims to push the envelope in unsupervised speech modelling, by taking the radical stance of trying to learn the full speech processing stack without any textual resources. Here, we reopen two previous benchmarks with a focus on discovering discrete representations from raw audio at two linguistic levels. The first focuses on the phonological or sub-word level. The goal is to learn discrete (low bitrate) speech units, which encode meaningful linguistic invariants, and which are useful for doing speech synthesis. This is a reopening of the 2019 "TTS without T" ZeroSpeech benchmark [3] (track 1b in Figure 1). The second focuses on the word level. The goal is to discover word-like units for the purpose of segmenting continuous speech. This is a reopening of the 2017 "spoken term discovery" ZeroSpeech Benchmark [2] (track 2 in Figure 1). As before, we rely exclusively on freely accessible software and data sets.

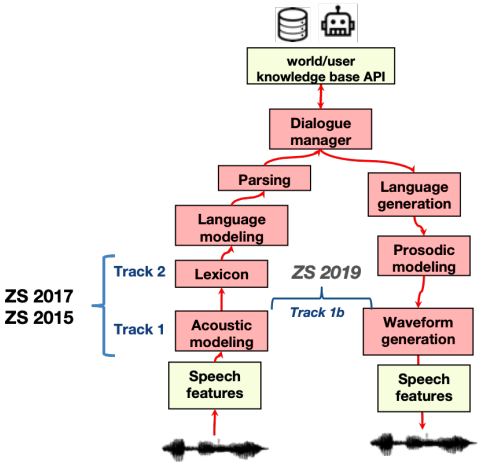

Figure 1: *Schematic diagram of the Zero Resource challenge series. The long term aim is to learn an entire spoken dialogue stack without recourse to any textual resources, mimicking the way human children learn their native languages.*

Discrete units, such as words and phonemes, form the basis of every modern speech technology system at some level. One useful feature of discrete representations is that they remove linguistically irrelevant information from the signal, and represent continuous speech in a highly compact format. For example, [4] perform unsupervised representation learning, and show that, up to a certain point, discrete representations are more useful than continuous ones as the input for training a phone recognizer. Here, we ask participants to discover their own discrete units and analyze them in terms of how well they capture relevant linguistic contrasts, as indicated by the gold phoneme- and word-level transcriptions.

[1]Detailed results of all metrics, and audio samples for unit discovery/synthesis systems, are provided on the leaderboard at `http://www.zerospeech.com/`.

http://dx.doi.org/10.21437/Interspeech.2020-2743

2. Data sets, metrics and baselines

2.1. Unsupervised unit discovery for speech synthesis

Task. The problem of learning speech units useful for doing speech synthesis can be seen in terms of an encoder–decoder architecture. The encoder takes as input raw audio and turns it into a sequence of speaker-invariant discrete units ("pseudo-text"). The decoder takes these units as input and generates a waveform corresponding to the same linguistic content uttered in another voice. Requiring synthesis is in a new voice allows us to exclude trivial solutions where the audio is returned unchanged. We measure the **synthesis quality** of the output, as well as the **unit quality** and the **bitrate** of the pseudo-text.

Data sets. We use the same setting and datasets as in [3], with two languages, the development language (English) and the surprise language (Indonesian). Participants are instructed to treat these languages as low-resource and refrain from using other labelled data. Metrics are provided only for the development language, and results for the surprise language must be submitted through the challenge website to be evaluated (maximum two submissions per research group). Three unlabelled data sets are provided for each language.The *Train Voice* data set contains either one talker (surprise) or two (development), and is intended for building an acoustic model of the target voice for speech synthesis (between 1h30 and 2h40 of data per voice). The *Train Unit Discovery* data set contains read speech from multiple speakers, with around ten minutes of speech from each speaker, for a total of 15h in each language. These are intended for the discovery of speaker-independent acoustic units. The *Test* data set contains new utterances from unseen speakers.

Synthesis quality is measured using human evaluation, taking three measures. To evaluate the *comprehensibility* of the synthesis, the evaluators were asked to orthographically transcribe the synthesized sentence. Each transcription was compared with the gold transcription using the Levenshtein distance, yielding a character error rate (CER). The overall *naturalness* of the synthesis was assessed on a 1 to 5 scale, yielding a mean opinion score (*MOS*), where 5 is the most natural. Finally, the *similarity* of the voice to the target voice was assessed using a 1 to 5 scale, with each synthesized sentence compared to a reference sentence uttered by the target voice, with 5 being the most similar to the target. Each evaluator performed the evaluation tasks in the same order (comprehensibility, naturalness, similarity), with the overall evaluation lasting about one hour. We recruited evaluators for English using the Amazon Mechanical Turk platform and, for Indonesian, through universities and research institutes in Indonesia. All were paid the equivalent of 10 USD. Catch trials (novel natural recordings by the target voice) ensured that participants were on task: only data from participants with $<$0.80 CER on catch trials was retained (English: 35; Indonesian: 68).

Embedding bitrate and quality. Participants submit encodings for each test file. The submitted encodings are sequences of vectors. These vectors are assumed to be quantized. We calculate the *bitrate* of the encoding by, first, constructing a dictionary of all distinct vectors over the entire test set. The test set is seen as a sequence $U = [s_1, ..., s_n]$ of n symbols. The bitrate is calculated as $n \sum_{i=1}^{n} \frac{p(s_i) log_2 p(s_i)}{D}$, where $p(s_i)$ is the relative frequency of symbol s_i in U, and D the total duration of U in seconds.The *unit quality* is evaluated with the *ABX phone discriminability score*, as in previous Zero Resource challenges [5, 2]. The ABX discriminability, for example, between [aba] and [apa], is defined as the probability that the representations of A and X are more similar than representations of B and X, over all triplets of tokens such that A and X are tokens of [aba], and B a token of [apa] (or vice versa), and such that X is uttered by a different speaker than A and B. The global ABX phone discriminability score aggregates over the entire set of minimal triphone pairs such as [aba]–[apa] to be found in the test set. The choice of the appropriate distance measure is up to the researcher. As in previous challenges, we provide a default distance, the average frame-wise angle (arc cosine of the normalized dot product) between the embeddings of the tokens along a DTW-realigned path, and also make available an equivalent distance making use of frame-wise symmetrised KL-divergences, rather than angles, as well as a Levenshtein (edit) distance measure. We cite ABX scores as error rates (0% for the gold transcription, 50% being chance). Each of the items compared in the ABX task is a triphone ([izi]-[idi], and so on), extracted from the test corpus. Each triphone item is a short chunk of extracted audio, to be decoded by the systems.[2]

Toplines and baselines. A baseline system is provided, consisting of a pipeline with a nonparametric Bayesian acoustic unit discovery system [6, 7], and a parametric speech synthesizer based on Merlin [8]. As linguistic features, we use contextual information (leading and preceding phones, number of preceding and following phones in current sentence), but no features related to prosody, articulatory features (vowel, nasal, and so on), or part-of-speech (noun, verb, adjective, and so on). The baseline system is made available in a container. A supervised topline system is also provided, consisting of a phone recognizer trained using Kaldi [9] on the original transcriptions. The acoustic model is a tristate triphone model with 15000 Gaussian mixtures. The language model is a trigram phone-level language model.[3] Output is piped to the TTS system, which is also trained on the gold labels.

Table 1: *Submissions to the **unit discovery/synthesis** track.*

	Encoder/Decoder	**Generation**
MC [10] (Sheffield)	Disentangled discrete AEs	Wavenet
TM [11] (Kyoto)	ABCD-VAE	Neural source-filter
BN [12] (SU)	VQVAE (1), VQCPC (2)	WaveRNN
PL [13] (Nagoya)	CycleVQVAE (1), Cycle-VAE (2)	Wavenet
BY (Brno)	Subspace HMM + AUD + Baseline	Baseline
MK [14] (IIT)	CV, VC transients	Waveglow
AT [15] (NAIST)	VQVAE + transformer	Griffin-Lim (2)
BG [16] (Boğaziçi)	Correspondence rec. sparse AE	Baseline
WH (Tokyo IT)	Hierarchical VQ-VAE	MelGAN

[2]This differs from previous challenges. In previous challenges, longer audio files were provided for decoding, from which the representations of triphones were extracted after the fact using time stamps. In the 2019/2020 edition, triphones are pre-extracted, to allow for systems without fixed frame rates.

[3]A word-level language model gives better performance, but we use a phone-level language model in the interest of giving a fair comparison with the subword unit discovery systems asked for in the challenge.

2.2. Spoken term discovery & segmentation

Task. The goal of spoken term discovery is to find words in the speech stream—just as the infant learns the words of its language by listening. The input is a series of speech features. The output is a set of boundaries delimiting the start and end of proposed word tokens discovered in the speech, and category labels indicating proposed word types. These boundaries may, but need not, constitute an exhaustive parse of the speech. The evaluation we apply is a set of scores measuring different aspects of the alignment with the words in the gold-standard transcription. As is customary in the field of word segmentation, we do not provide a separate test set for this track; we rely on the surprise languages to assess possible hyperparameter overfitting. Two submissions per research group are allowed.

Data sets. The *development data* and *surprise data* are the same as in [2] (see [17, 18]). The development data consists of corpora from three languages (English, French and Mandarin). Each corpus comes with software that performs the evaluation. Challenge participants are encouraged to use these resources to tune their hyperparameters using a cross-validation approach to maximize generalizability. The participants then must submit their systems and their output on all the data sets for independent evaluation (run automatically upon submission). The surprise data consists of corpora from two additional languages (German and Wolof), which are provided with no additional resources.

The amount of data in the training part of the development data sets varies from 2.5 to 45 hours, to ensure that systems can work both with limited data and with larger data sets. The statistics of the two surprise languages fall between these two extremes. The distribution of speakers in the training sets is shaped to reflect what is typically found in natural language acquisition settings: there is a "family"—a small number of speakers (between four and ten) who make up a large proportion of the total speech—and a set of "outsiders"—a larger number of speakers that each appear in smaller proportions (ten minutes each). The test sets consist of many short files, are organized into subsets of differing length (1s, 10s and 120s).

The English and French corpora were taken from LibriVox audio books[4] and phone force-aligned using Kaldi [9]. The Mandarin corpus is described in [19], force-aligned using Kaldi. The German corpus was taken from LibriVox and force-aligned using Kaldi as well. The Wolof corpus is described in [20].

Evaluation metrics. "Spoken term discovery" is a complex task with a number of sub-goals, which can be evaluated separately. The first sub-goal is to do good *matching:* deciding whether any two given speech fragments are instances of the same sequence of phoneme, and attempting to find as many matches as possible. The *quality of matches* is evaluated based on how similar fragments matched by the system are—we use the average normalized edit distance (**NED**) between the gold phoneme sequences, over all pairs matched by the system—the *quantity of matches* can be evaluated by measuring the proportion of the corpus covered by matched pairs (**coverage**).

The second sub-goal is to construct a lexicon. This amounts to *clustering* the discovered matched pairs. The *intrinsic quality* of the lexicon is evaluated based on how consistent the items clustered together are with regard to the sequences of gold phonemes they correspond to. The **Grouping** scores (precision, recall and F-score) evaluate the purity and inverse fragmentation of the clusters in a pairwise fashion (see [2] for a formal definition). The *extrinsic quality* can be measured with respect to how well the clusters match the gold-standard lexicon of word types. **Type** scores (precision, recall and F-score) measure the correspondence between the discovered clusters and the gold lexicon. Type precision is the probability that discovered types belong to the gold set of types (real words), type recall is the probability that gold types are discovered. We restrict both sets to words between three and twenty phones long.

The third sub-goal is to do *accurate word segmentation.* The **Token** scores (precision, recall and F-scores) evaluate the quality of the discovered fragment tokens compared to the gold tokens, and the **Boundary** scores (precision, recall and F-scores) the quality of the discovered boundaries.

By setting out three different types of criteria, the intention is to be open to various types of "spoken term discovery" systems, all of which in some sense "find words." The result is that we do three (non-independent) types of evaluations. All of these evaluations are done at the level of the phonemes: using the aligned phoneme transcription, we convert any discovered fragment of speech into its transcribed string. If the left or right edge of the fragment contains part of a phoneme, that phoneme is included in the transcription if it corresponds to more than 30ms or more than 50% of its duration.

Baselines and toplines. The baseline was computed using [21], which does pair-matching using locally sensitive hashing applied to PLP features and then groups pairs using graph clustering. The parameters stayed the same across all languages, except that the dynamic time warping threshhold was increased for Mandarin (to 0.90, rather than 0.88), in order to obtain a NED value similar to that of other languages. The topline system was an exhaustive-parsing word segmentation model based on the textual transcriptions (a unigram grammar trained in the adaptor grammar framework: [22]).

3. Models and selected results

3.1. Unsupervised unit discovery for speech synthesis

Sixteen systems from nine teams were submitted, summarized in Table 1. Two systems, **AT-1** and **BG**, are excluded from analysis of the synthesis evaluation due to declared issues with the submissions. Relatively few systems were submitted in the "low bitrate" range (near the bitrate of the annotation). Nevertheless, the systems submitted this year, which have shifted towards higher bitrates and end-to-end systems mostly based on discrete autoencoders, have all done more with less. Figure 2a shows (for the surprise language) the improvements in embeddings with respect to the previous year: the edge of the grey zone shows the empirical tradeoff previously observed between **unit quality** and **bitrate**. Many of this year's systems improve reach lower ABX error rates at a given bitrate. Figure 2b shows that improvements have also been made in **decoding,** with overall more comprehensible synthesis, regardless of unit quality (the 2019 systems are represented by the dotted line of best fit, while the solid line is fit through the current submissions). And, while the comprehensibility measure is largely correlated with the overall synthesis naturalness evaluations (MOS), Figure 2c shows that certain systems are reported to sound particularly natural, beyond just their comprehensibility (notably the two **MK** systems). This is presumably due to a improvement in **waveform generation.** Figure 2d shows the combined effect of these improvements in **unit quality**, **decoding,** and **waveform generation,** showing major improvements on the tradeoff between synthesis quality and bitrate.

[4] http://librivox.org/

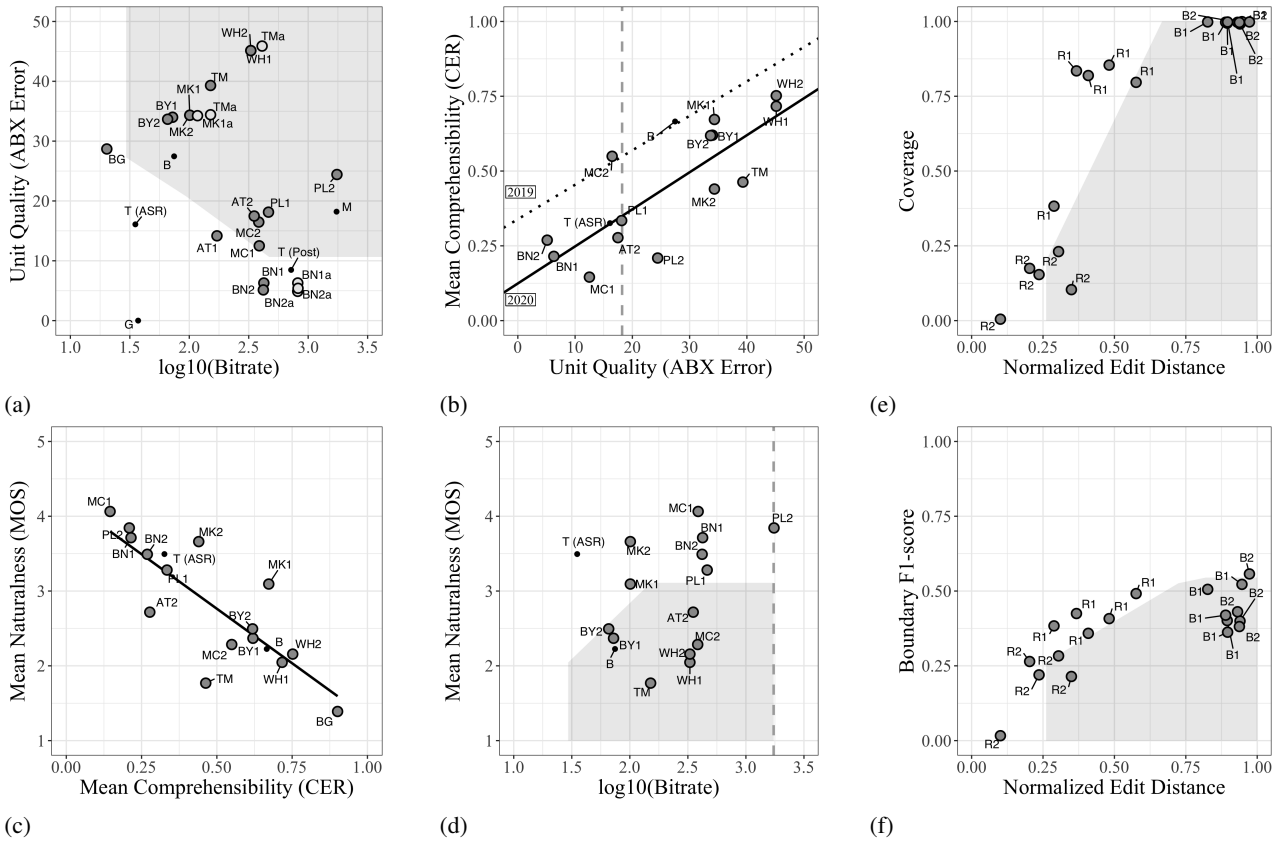

Figure 2: ***(a)*** *ABX error (lower better) as function of bitrate for* ***unit discovery/synthesis****.* ***(b)*** *Character error rate (CER: lower is better) as a function of ABX error. Vertical dashed line: ABX error for MFCCs. Sloped lines are linear regressions for 2019 submissions (dotted) and for 2020 submissions (solid), showing global increase in decoding quality.* ***(c)*** *Mean opinion score (MOS: higher is better) as a function of CER. Line is linear regression.* ***(d)*** *MOS as function of bitrate. Vertical dashed line is MFCC bitrate.* ***Unit discovery/synthesis*** *results presented on surprise language only. Reference scores plotted as* ***G*** *for gold transcriptions;* ***M*** *for MFCC features;* ***B*** *for baseline system;* ***T (Post)*** *for posteriorgrams from the topline system; and* ***T (ASR)*** *for discrete decoding from the topline. Edge of grey regions in* ***(a)*** *and* ***(d)*** *represents 2019 state of the art on the tradeoff. Labels are 2020 submissions.* ***(e)****. Coverage (higher is better) as a function of normalized edit distance (NED: lower is better) for* ***spoken term discovery/segmentation*** *submissions.* ***(f)*** *Boundary F-score (higher is better) as a function of NED for submissions. Edge of grey regions in* ***(e)*** *and* ***(f)*** *represents 2017 state of the art on the tradeoff, labels are 2020 submissions, and multiple points per system are different languages. Clustering-oriented algorithms have low NED, while segmentation-oriented algorithms have high coverage and boundary F-scores.*

3.2. Spoken term discovery & segmentation

Two teams, indicated in Figure 2 as **B** [23] (JHU) and **R** [24] (Tampere), submitted two systems each. The edge of the grey region in Figure 2e shows the empirical tradeoff previously observed between having **high quality matching** (low NED) and **exhaustively analysing** the corpus (high coverage). Systems **R1** and **R2**, which employ probabilistic dynamic time warping, both clearly improve on the tradeoff, with **R1** privileging exhaustiveness and **R2** match quality. Figure 2f shows the empirical tradeoff between high quality matching and **accurate word segmentation**. Systems **R1** and **R2** again show improvement. Systems **B1** and **B2**, which use self-expressing autoencoders to improve frame representations before segmenting and clustering, show higher boundary F-scores, comparable to the previous state of the art for systems privileging segmentation.

4. Conclusion

Major advances have been made towards unsupervised unit discovery for speech synthesis, at all levels—better units, better decoding architectures, and better waveform generation. The best discrete codes, however, are still an order of magnitude more detailed than the phonemic representation. The supervised topline system demonstrates the possibility of a low bitrate code which is also of high quality. The challenge is to find such a high-quality low-bitrate phoneme-like representation in an unsupervised fashion. Nevertheless, some higher-bitrate codes may yet be useful, and good enough, to be used in language modelling. We will explore this in upcoming challenges. Regarding spoken term discovery and segmentation, progress was made in this challenging dual task, with improved clusters and improved coverage. Clustering-oriented algorithms represent the best current tradeoff, but another potential path forward is to bring segmentation-oriented systems towards better clusters.

5. Acknowledgements

Funded by the ANR (ANR-17-EURE-0017 FRONTCOG, ANR-10-IDEX-0001-02 PSL*, ANR-18-IDEX-0001 U de Paris, ANR-19-P3IA-0001 PRAIRIE 3IA Institute, ANR-17-CE28-0009 GEOMPHON, ANR-10-LABX-0083 EFL), CIFAR LMB, and a research gift by Facebook.

6. References

[1] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results." in *SLTU*, 2016, pp. 67–72.

[2] E. Dunbar, X. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," *arXiv:1712.04313*.

[3] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black *et al.*, "The zero resource speech challenge 2019: Tts without t," *arXiv preprint arXiv:1904.11469*, 2019.

[4] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS

[5] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," *Procedia Computer Science: Proceedings of SLTU 2016*, vol. 81, pp. 67–72, 2016.

[6] L. Ondel, L. Burget, and J. Cernocký, "Variational inference for acoustic unit discovery," in *SLTU*, ser. Procedia Computer Science, vol. 81. Elsevier, 2016, pp. 80–86.

[7] L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. Khudanpur, "Bayesian models for unit discovery on a very low resource language," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5939–5943.

[8] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Speech Synthesis Workshop*. ISCA, 2016, pp. 202–207.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[10] M. Chen and T. Hain, "Unsupervised acoustic unit representation learning for voice conversion using wavenet auto-encoders," in *Proceedings of INTERSPEECH 2020*, 2020.

[11] T. Morita and H. Koda, "Exploring TTS without T Using Biologically/Psychologically Motivated Neural Network Modules (ZeroSpeech 2020)," in *Proceedings of INTERSPEECH 2020*, 2020.

[12] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," in *Proceedings of INTERSPEECH 2020*, 2020.

[13] P. Lumban Tobing, T. Hayashi, Y.-C. Wu, K. Kobayashi, and T. Toda, "Cyclic spectral modeling for unsupervised unit discovery into voice conversion with excitation and waveform modelingh," in *Proceedings of INTERSPEECH 2020*, 2020.

[14] K. Pandia, A. Prakash, M. R. Kumar, and H. Murthy, "Exploration of End-to-end Synthesisers for Zero Resource Speech Challenge 2020," in *Proceedings of INTERSPEECH 2020*, 2020.

[15] A. Tjandra, S. Sakti, and S. Nakamura, "Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge," in *Proceedings of INTERSPEECH 2020*, 2020.

[16] B. Gündoğdu, B. Yusuf, M. Yesilbursa, and M. Saraclar, "Vector quantized temporally-aware correspondence sparse autoencoders for zero-resource acoustic unit discovery," in *Proceedings of INTERSPEECH 2020*, 2020.

[17] 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.

[18] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *ASRU 2017*, 2017, pp. 301–308.

[19] D. Wang, X. Zhang, and Z. Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[20] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui, "Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof," *LREC*, 2016.

[21] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 401–406.

[22] S. Goldwater, T. L. Griffiths, and M. Johnson, "A bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[23] S. Bhati, J. Villalba, P. Żelasko, and N. Dehak, "Self-expressing autoencoders for unsupervised spoken term discovery," in *Proceedings of INTERSPEECH 2020*, 2020.

[24] O. Räsänen and M. A. Cruz Blandón, "Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics," in *Proceedings of INTERSPEECH 2020*, 2020.