Lifelong Learning MOS Prediction for Synthetic Speech Quality Evaluation

Félix Saget, Meysam Shamsi, Marie Tahon

Interspeech 2024
1-5 September 2024, Kos, Greece

Lifelong Learning MOS Prediction for Synthetic Speech Quality Evaluation

Félix Saget[1], *Meysam Shamsi*[1], *Marie Tahon*[1]

[1]LIUM, Le Mans University, France

Felix.Saget@univ-lemans.fr, Meysam.Shamsi@univ-lemans.fr, Marie.Tahon@univ-lemans.fr

Abstract

Mean Opinion Score (MOS) has been a long-standing standard for perceptive evaluation of quality of speech synthesis models; however, this criterion is hardly reproducible, and costly. Automatic, neural MOS predictors have emerged as a solution to the objective assessment of synthetic speech. These predictors are trained once on data collected from past listening tests, and thus may suffer from adaptation to new technology breakthrough in speech synthesis. In this study, we investigate the applicability of lifelong learning for MOS predictors, where the training samples would be fed to the model in the chronological order. A sequential lifelong mode and a cumulative lifelong mode have been compared with traditional batch training using the BVCC and Blizzard Challenge datasets. The experiments show the advantages of lifelong learning in cross-corpus evaluation as well as in a constrained data availability scenario.

Index Terms: speech synthesis, lifelong learning, speech quality evaluation, MOS prediction

1. Introduction

Over the recent decades, speech synthesis technologies have undergone significant evolution, advancing from unit selection and Hidden Markov Models to deep neural networks (see Figure 1). Although each new approach faced its own unique challenges, the synthesis quality of state-of-the-art Text-to-Speech (TTS) and Voice Conversion (VC) systems has been steadily going upwards, to a point where the latest state-of-the-art neural systems are claimed to be indistinguishable from natural speech [1, 2].

Synthetic speech quality is a multi-aspect concept which is hard to define and even harder to assess. Three criteria are typically considered: intelligibility (is the lexical content understandable?), naturalness (does the sample sound "human"?), and expressivity (does the sample convey dynamics, intent, emotion?). This study focuses on the evaluation of naturalness. The most reliable approach for assessing the synthetic quality of a TTS system involves conducting a perceptual test with human listeners, which are instructed to give a subjective estimation of a certain aspect of quality. This evaluation method comes with certain limitations and drawbacks [3, 4]: its outcome may exhibit variability [1] due to the inconsistencies in the testing environment, the nature of listeners, and the formulation of instructions from one test to another.

The Mean Opinion Score (MOS) [5] stands as the prevalent non-intrusive (without reference speech) method for evaluating synthetic speech, involving the assignment of a quality score to a given input signal. Other evaluation approaches, such as AB comparison test or MUSHRA test [6] do exist; however, despite facing criticism [4], MOS continues to be the most widely used perceptual evaluation method in the community, as a more efficient alternative is yet to emerge.

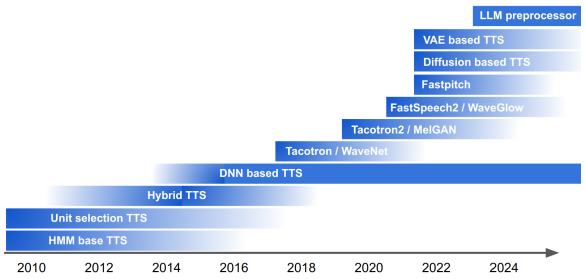

Figure 1: *A timeline of TTS models in the last decade.*

The MOS provided by human listeners can serve as label of training samples for a quality predictor system, to automate the process of synthetic quality evaluation. Recent research has exhibited a growing interest in this measurement, with the introduction of systems such as MOSA-Net [7], LDNet [8], and SSL-MOS [9]. Noteworthy initiatives like the VoiceMOS Challenge [10, 11] have established a comparative platform, featuring a standardized evaluation protocol for predictive models, and a publicly available dataset.

By focusing on the MOS predictor as a regressor model, any MOS evaluation of the output from TTS and Voice Conversion systems can serve as a training set. With the proliferation of various types of TTS systems, the accumulation of data through MOS evaluations has led to the creation of diverse datasets [12, 13]. However, the combination of human answers from different tests may prove inefficient, especially when dealing with distinct domains such as different languages, applications, or TTS models. Notably, evaluations conducted through the years can be influenced by contextual factors and listeners' expectations of TTS quality. As reported in [13, 14], the quality score of synthetic speech can be relative to available resources and the time of evaluation.

The primary objective of this study is to incorporate the yearly evolution of TTS systems into the training of the MOS predictor. To the best of our knowledge, this is the first study to introduce a lifelong learning approach [15, 16, 17, 18], also known as continual learning, for MOS prediction. The training data, consisting of synthetic signals and corresponding human-sourced MOS scores, would be sequentially fed into the MOS predictor model based on chronological order. The MOS prediction performance of this approach will be compared to the conventional method of providing training data as a whole batch. The motivation behind adopting a lifelong learning strategy is rooted in the availability of training data spanning differ-

10.21437/Interspeech.2024-959

ent periods. The main questions to be addressed in this study are as follows:

- When new data is collected, is it essential to re-train a model, or is fine-tuning the last checkpoint sufficient?
- Does the cost of re-evaluating old synthetic speech in the same period of time as training MOS predictors offer any advantages?
- What is the impact of adhering to the timeline of the data on the generalization performance of a MOS predictor?
- How do limitations in resources, including computational constraints and data volume, alter the results when comparing the classical approach to the lifelong approach?

The study concentrates on assessing the impact of incrementally nurturing the training data for the MOS predictor model, particularly for utterance-level quality evaluation. The underlying concept is that the training process can mirror the evolution of TTS models with respect to quality, drawing parallels to the notion of curriculum learning [19], where the order at which training samples are fed impacts the final results.

2. Method

To explore the impact of training synthetic quality predictors in line with the historical availability of data, a temporal dataset is required. The subsequent sections provide details on the datasets, the baseline system used as the MOS predictor, and the experimental protocol.

2.1. Data

The Blizzard Challenge [20] has been a long-standing benchmark for state-of-the-art TTS evaluation. Organizers have conducted large-scale listening tests on synthetic speech samples provided by participants using contemporary systems.

The collection of perceptual scores over time has created an archive of the evolution of TTS systems, spanning the transition from unit selection to DNN-based TTS. With each edition of the Challenge focusing on different aspects of speech synthesis (language, amount of data...), and the heterogeneity in the models used and their specific artifacts, using traditional machine learning schemes on this historical data seems unreasonable. This motivates our investigation of lifelong learning.

In preparation for the VoiceMOS Challenge 2022 [10], synthetic English speech samples from Blizzard Challenges 2011-2016 were collected. This data was enriched with synthetic speech excerpts from Voice Conversion Challenges 2016, 2018 and 2020, and additional samples produced with various ESPnet TTS systems [21], amounting to 187 systems in total (including natural speech). These samples were submitted to a new, large-scale listening test in 2021: the human answers were gathered into a dataset called *BVCC*.

However, we argue that human answers obtained from a single listening test (at a single point in time) are not equivalent to those gathered through listening tests conducted in the past: a listener from 2008 might not have the same expectation of synthetic speech quality as a listener from 2021 [14], nor will a future listener. This passive bias may affect the listener's opinion. Consequently, we separately collect listening test answers, in terms of naturalness, from the Blizzard Challenges 2008-2013, on primary English tasks (EH1) only. This dataset will be referred to as *BC* in the following. Table 1 displays the number of samples in *BVCC* and *BC* dataset with the different partitioning of *train/val/test* sets.

Table 1: *Number of samples in two datasets (BC and BVCC) and train/validation/test partitions*

Blizzard Challenge data (*BC*)									
	Batch mode			Lifelong mode			C. Lifelong mode		
Year	train	val	test	train	val	test	train	val	test
2008				706	176		706	176	
2009				545	136		1251	312	
2010	2529	631	-	518	130	-	1769	442	-
2011				406	101		2175	543	
2012				354	88		2529	631	
2013	-	-	563	-	-	563	-	-	563
VoiceMOS Challenge data (BVCC)									
2008				638	160		638	160	
2009				546	137		1184	297	
2010				547	137		1731	434	
2011	4374	1096	-	395	99	-	2126	533	-
2013				334	84		2460	617	
2016				1124	281		3584	898	
2018				790	198		4374	1096	
2020	-	-	1254	-	-	1254	-	-	1254

2.2. Experiment

2.2.1. Choice of MOS prediction system

MOS prediction systems are typically prone to bad generalization, given the context heterogeneity from one listening test to another [9]. State-of-the-art systems thus take advantage of large speech feature extractors such as HuBERT [22] or wav2vec2 [23] to gain in generalization ability, and can even reach good performance on zero-shot prediction on unseen systems [11].

In order to select a system for our experiments, we ran a comparison of the three baseline models of the VoiceMOS Challenge 2022 [10]. SSL-MOS[1] [9] predicts MOS naturalness by appending a single linear layer to wav2vec2.0-base. MOSA-Net[2] [7] uses spectral features, waveform, and HuBERT embeddings. Finally, LDNet[3] [8] models a listener-dependent MOS score directly from spectrograms (no large feature extractor involved).

We compared the three architectures by retraining them from scratch on the *BVCC* dataset, measuring correlation to ground truth (Spearman Ranking Correlation Coefficient, following [10]) as well as training times. The official implementations were used. The three models achieved comparable accuracy, however, SSL-MOS has a much lower computational time for training and inference. Consequently, we use this architecture for our experiments.

2.2.2. Lifelong learning for MOS prediction

The typical training protocol for deep learning models consists of iteratively feeding the entire dataset $\mathcal{X}$ in a random order, with a stopping criterion after a given number of epochs. We refer to this approach as *Batch mode* in the following.

In our specific context, the dataset $\mathcal{X}$ is collected and annotated over multiple time periods: $\mathcal{X} = \mathcal{X}_{t_0} \cup \mathcal{X}_{t_1} \cup ... \cup \mathcal{X}_{t_n}$. We are looking to train a model on this temporal data which is able to predict the synthetic speech quality of newly collected data $\mathcal{X}_{t_{n+1}}$ (test set) as faithfully as possible to the opinion of a real

[1] github.com/nii-yamagishilab/mos-finetune-ssl
[2] github.com/dhimasryan/MOSA-Net-Cross-Domain
[3] github.com/unilight/LDNet

human listener. We assume that the implicit chronological information could help the model gain a better comprehension of the evolution of naturalness perception over time, and could be leveraged to improve prediction accuracy on new synthetic samples of unseen quality. In the lifelong learning training protocol, we begin by training a model on $\mathcal{X}_{t_0}$, then resume training on $\mathcal{X}_{t_1}$ only, until all data has been seen (t_n). The data corresponding to each time period is fed sequentially to the model, without preserving old data. This is referred to as *Lifelong mode* in the following. Additionally, we consider the same training protocol, but by keeping data from all previous time periods $\bigcup_{i=1}^{n} \mathcal{X}_{t_i}$. We call this protocol Cumulative Lifelong mode (*C. Lifelong mode*).

By comparing these three modes of learning, we study the utility of aged training samples and the benefit of recycling models trained on previous data. In *Batch mode*, we do not recycle a trained model and use all data to train from scratch, which gives same importance to all periods. In *Lifelong mode*, we recycle a model trained on previous data, and task the model with focusing only on new data and forgetting aged data [16]. In *C. Lifelong mode*, we recycle a model trained on previous data and let it access all available data up to time t_i.

For each training protocol, we randomly split the available data at each time step (year) into training and validation sets ($80/20\%$), as shown in Table 1. All models are trained for maximum 100 epochs, with early stopping of 5 epochs on validation, using Stochastic Gradient Descent (SGD). The SGD momentum is set to 0.9, the learning rate to $1\mathrm{e}^{-5}$, and the shuffled batch size to 4. Following [9], for a given audio sample, the Mean Absolute Error (MAE) between the predicted value and its ground truth label is used as the loss function. Finally, we evaluate MOS prediction on the last available year (test) of *BC* and *BVCC*.

2.2.3. Cross corpus evaluation

In order to measure the generalization of our MOS predictors to unseen annotations, we evaluate the model trained on *BC-train* on *BVCC-test* and vice-versa. It should be reminded that part of *BC-test* samples were annotated in BVCC: we did not discard those samples.

Furthermore, we also evaluate models trained on *BC* and *BVCC*, on the test split (3001 files) of the *SOMOS* dataset [12]. *SOMOS* is a crowdsourcing evaluation of MOS naturalness of 200 English TTS systems collected in 2022. In order to keep the same sampling rate between datasets, the samples of *SOMOS* are downsampled from 24kHz to 16kHz, assuming sufficient perceptual consistency. We note that the perceptual impact of downsampling on MOS is under-studied; we leave any investigation of this subject to future works.

2.2.4. Limited resources

The ultimate goal of a MOS prediction system is to mitigate the need for human MOS answers. We investigate the performance of models trained with our three protocols in conditions where the annotation budget is limited. To simulate this situation, the available data is randomly sampled down to 50% and 25%. Test sets are left unchanged from Table 1.

We also introduce an additional version of *Lifelong mode* where instead of disposing of all the old data, time periods are considered in a "sliding window" fashion. So as new data is added to the training data, the most ancient data is removed. In this case, the model have access to the two most recent years in the lifelong mode ($\mathcal{X}_{t_{i-1}} \cup \mathcal{X}_{t_i}$).

3. Results

In line with the recommendation of [11], which underscores the significance of quality rankings and the challenges associated with interpreting MOS values [24], we choose to compare performances using SRCC. Table 2 presents the SRCC of the MOS predictor trained on *BC* and *BVCC* datasets, and evaluated on *BC*, *BVCC* and *SOMOS* dataset.

3.1. Fine-tuning or retraining?

Table 2 showcases how *Lifelong mode* does not improve SRCC in comparison to *Batch mode*, when evaluated on the test split of the dataset it was trained on. However, there is no significant difference in SRCC results between *Batch* and *C. Lifelong mode* across both datasets.

Through the examination of the evolution of training with *Lifelong* and *C. Lifelong mode*, a degradation in performance on the test set has been noted when incorporating samples from 2013 in *BVCC* and 2012 in *BC*. This phenomenon is illustrated in Figure 2. The rationale behind this observation may stem from the distinctive characteristics of samples from these two years compared to the rest of the dataset: indeed, 2012 and 2013 mark the emergence of hybrid TTS (HMM + DNN or unit selection + DNN), resulting in synthetic speech with unique attributes. Additionally, as reported by [14], there was an excessive control of prosody in the 2013 Blizzard Challenge dataset, and TTS systems participating in the 2013 Blizzard challenge benefited from a substantial volume of training data (300h), leading to differences in synthetic quality compared to resources from other years.

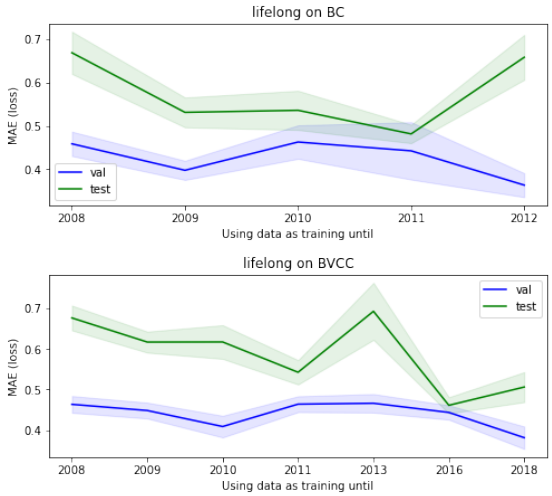

Figure 2: *Evolution of MAE as loss function on test and validation in Lifelong mode. The val set corresponds to current year of training and test set corresponds to last available year.*

Conducting an experiment on *Lifelong mode* without incorporating data from these specific years (2012 in *BC* and 2013 in *BVCC*) showed no improvement in prediction accuracy on the test sets, implying this data is not useful to the learning process. The lifelong learning approach enabled this investigation, revealing the distinct nature of samples in certain years.

As a prospective application, the utility of conducting a perceptual test to annotate new data can be validated through the evaluation of the utility of a small subset.

3.2. Cross dataset evaluation

Table 2 shows a notable limitation in the generalization ability of the MOS predictor across datasets. This observation aligns with reported performances in [9]. A limitation of the classical *Batch mode* in generalization can be highlighted by examining the evaluation results on *BC-test*. A comparison between *Batch mode* training with the *BC-train* and *Batch mode* training with the *BVCC-train*, which shares common synthetic signals with the *BC-test*, exposes this limitation. Training with the *BVCC* dataset demonstrates lower performance (SRCC=0.74) compared to the *BC* dataset (SRCC=0.84), despite the latter containing a smaller number of samples.

However, the generalization ability of *Lifelong mode* (particularly *C. Lifelong mode*) is significantly higher than *Batch mode* for most of the cases, which demonstrates an advantage of this approach. For instance, examining the last column of Table 2 reveals the advantages of training in *C. Lifelong mode* when evaluating on *SOMOS-test*.

Another notable observation is that using the *BC-train* is more efficient (with a lower number of samples and higher performance) compared to using *BVCC-train* in cross-dataset, when the test set is *SOMOS*. Some may argue that the lower performance may be attributed to the presence of Voice Conversion samples in *BVCC*. To test this hypothesis, utilizing only common data (2008-2011) for both *BC* and *BVCC*, with an equal number of training samples and excluding voice conversion samples, indeed confirms the advantage of the *BC* dataset over *BVCC*. The recent observation raises doubts about the necessity of re-annotation of synthetic speech on the day of training MOS predictors (the case of *BVCC*), especially when it incurs costs and leads to a degradation in performance.

Table 2: *SRCC results of training on different datasets and modes. Confidence intervals are calculated on 10 runs.*

Training		Test		
Data	Mode	BC (2013)	BVCC(2020)	SOMOS
BC <2013	Batch	0.837±0.008	0.721±0.060	0.490±0.039
	Lifelong	0.811±0.010	0.737±0.004	0.484±0.019
	C. Lifelong	0.835±0.004	0.734±0.003	**0.555±0.013**
BVCC <2020	Batch	0.744±0.110	0.831±0.006	0.337±0.120
	Lifelong	0.820±0.010	0.814±0.007	0.353±0.026
	C. Lifelong	**0.836±0.006**	0.811±0.009	0.367±0.020

3.3. Limited resources

Table 3 presents the performance of *Batch mode* training and *C. Lifelong mode* in situations where the annotation budget is constrained. A significant decline in performance on the *BC* dataset (total 2529 training samples) is evident when only 25% (or even 50%) of the samples are utilized, emphasizing the importance of utilizing all available samples in *Batch mode*. This performance degradation is less pronounced on the *BVCC* dataset (total 4374 training samples).

One advantage of the proposed *C. Lifelong mode* is its ability to maximize information extraction in scenarios with limited annotated samples. A notable contrast in performance between *Batch mode* training and *C. Lifelong mode* is observed, especially when the number of training samples is more restricted (utilizing only 25% of the *BC* dataset). Examining the performance of these two training modes on the *BVCC* dataset, it is evident that achieving a performance of SRCC=0.74 requires using 2187 samples in the *Batch mode*, while only 1094 samples are sufficient in *C. Lifelong mode*.

Table 3: *Comparing SRCC of Batch and C. Lifelong modes when applying a constraint on training set size.*

		25%	50%
BC	Batch	0.218±0.071	0.495±0.132
	C. Lifelong	**0.805±0.011**	**0.818±0.008**
BVCC	Batch	0.571±0.141	0.740±0.029
	C. Lifelong	**0.745±0.011**	**0.775±0.010**

As explained in the experimental protocol (Section 2.2), the computation time required for the lifelong mode is higher than that for *Batch mode*. The number of training iterations (number of epochs multiplied by the number of batches) needed to obtain final checkpoints can indicate the necessary computational resources. The training iteration counted in *Batch* and *Lifelong mode* is almost the same. However, in *C. Lifelong mode*, this number is approximately two times higher on the *BC* dataset (and three times higher on the *BVCC* dataset). This disadvantage highlights the cost of *C. Lifelong mode*, or the increased opportunities for the model to be optimized during the training process. To strike a balance between performance and computational time for *Lifelong* and *C. Lifelong* approaches, the concept of sliding window lifelong (mentioned in Section 2.2.4) was tested. The results indicate that although the training iteration count falls somewhere between that of *Lifelong* and *C. Lifelong* mode, no significant difference of performance with lifelong was observed.

4. Conclusions

This paper has explored the use of lifelong learning for training MOS predictors for synthetic speech quality evaluation. The traditional *Batch mode* of training was compared with two lifelong learning modes: sequential *Lifelong mode* and *Cumulative Lifelong mode*. The experiments were conducted on the *BC* and *BVCC* datasets, with additional evaluations on the *SOMOS* dataset to assess generalization across datasets.

The findings suggest that fine-tuning a MOS predictor on new data, as opposed to retraining it from scratch, can yield benefits. While MOS predictors trained in *Batch mode* exhibited higher performance on the same dataset, the adoption of lifelong learning, especially *C. Lifelong mode*, demonstrated improved generalization across datasets. Additionally, in resource-constrained scenarios, such as reduced annotation budgets, *C. Lifelong mode* showed acceptable performance compared to the drastic degradation observed in the batch mode. As a perspective application, the proposed lifelong approach offers the opportunity for adaptation to new domains, such as new languages or more specific aspects of quality. The primary drawback of *C. Lifelong mode* remains its longer training time, which can be deemed a reasonable trade-off for a time-robust and reusable model.

Comparing the two available datasets of *BC* and *BVCC*, our experiments in predicting MOS scores on recent synthetic speech revealed that re-annotation of data is unnecessary, and preserving data from previous perceptual evaluations is more valuable.

5. Acknowledgements

This study has been realized under the PULSAR project supported by the Region of Pays de la Loire, France (grant agreement No 2022-09747) and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666 (Esperanto project).

6. References

[1] O. Perrotin, B. Stephenson, S. Gerber, and G. Bailly, "The Blizzard Challenge 2023," in *Proc. 18th Blizzard Challenge Workshop*, 2023, pp. 1–27.

[2] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Éva Székely, C. Tånnander, and J. Voße, "Speech synthesis evaluation — state-of-the-art assessment and suggestion for a novel research program," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 105–110.

[4] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the MOS pit: a critical analysis of MOS test methodology in TTS evaluation," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 41–47.

[5] "Mean opinion score terminology," *ITU-R Recommendation P.800.1*, 2016.

[6] "Method for the subjective assessment of intermediate quality level of audio systems," *ITU-R Recommendation ITU-R.BS.1534-3*, 2015.

[7] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.

[8] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDnet: unified listener dependent modeling in MOS prediction for synthetic speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 896–900.

[9] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[10] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," *arXiv preprint arXiv:2203.11389*, 2022.

[11] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2023: zero-shot subjective speech quality prediction for multiple domains," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.

[12] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "SOMOS: the Samsung Open MOS dataset for the evaluation of neural text-to-speech synthesis," in *Proc. Interspeech 2022*, 2022, pp. 2388–2392.

[13] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 183–188.

[14] S. Le Maguer, S. King, and N. Harte, "Back to the future: extending the Blizzard Challenge 2013." in *INTERSPEECH*, 2022, pp. 2378–2382.

[15] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: a review," *Neural networks*, vol. 113, pp. 54–71, 2019.

[16] Y. Prokopalo, S. Meignier, O. Galibert, L. Barrault, and A. Larcher, "Evaluation of lifelong learning systems," in *International Conference on Language Resources and Evaluation*, 2020.

[17] S. Sadhu and H. Hermansky, "Continual learning in automatic speech recognition." in *Interspeech*, 2020, pp. 1246–1250.

[18] M. Shamsi, A. Larcher, L. Barrault, S. Meignier, Y. Prokopalo, M. Tahon, A. Mehrish, S. Petitrenaud, O. Galibert, S. Gaist *et al.*, "Towards lifelong human assisted speaker diarization," *Computer Speech & Language*, vol. 77, p. 101437, 2023.

[19] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[20] The Blizzard Challenge. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge

[21] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESP-net: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[24] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the blizzard challenge 2007 listening test results," *The Blizzard Challenge Workshop (in Proc. SSW6)*, 2007.