TOWARDS LIFELONG LEARNING OF MULTILINGUAL TEXT-TO-SPEECH SYNTHESIS

Mu Yang[1], Shaojin Ding[2], Tianlong Chen[3], Tong Wang[4], Zhangyang Wang[3]

[1]University of Texas at Dallas, Richardson, Texas, USA
[2]Texas A&M University, College Station, Texas, USA
[3]University of Texas at Austin, Austin, Texas, USA
[4]University of Science and Technology of China, Hefei, Anhui, China

[1]mu.yang@utdallas.edu, [2]shjd@tamu.edu,
[3]{tianlong.chen, atlaswang}@utexas.edu, [4]wtwt1997@mail.ustc.edu.cn

ABSTRACT

This work presents a lifelong learning approach to train a multilingual Text-To-Speech (TTS) system, where each language was seen as an individual task and was learned sequentially and continually. It does not require pooled data from all languages altogether, and thus alleviates the storage and computation burden. One of the challenges of lifelong learning methods is "catastrophic forgetting": in TTS scenario it means that model performance quickly degrades on previous languages when adapted to a new language. We approach this problem via a data-replay-based lifelong learning method. We formulate the replay process as a supervised learning problem, and propose a simple yet effective dual-sampler framework to tackle the heavily language-imbalanced training samples. Through objective and subjective evaluations, we show that this supervised learning formulation outperforms other gradient-based and regularization-based lifelong learning methods, achieving 43% Mel-Cepstral Distortion reduction compared to a fine-tuning baseline.

Index Terms— Lifelong learning, Multilingual Text-To-Speech synthesis

1. INTRODUCTION

Multilingual text-to-speech (TTS) synthesis aims to synthesize speech of different languages given corresponding input texts. Conventional Multilingual TTS systems usually require an independent model for each language [1, 2]. More recently, end-to-end multilingual TTS system (i.e., one model for all languages) has been shown to achieve convincing performance [3, 4, 5]. These systems significantly reduce the deployment complexity, which is increasingly suitable for real-world use scenarios.

Current multilingual TTS systems typically require gathering data for all target languages before launching training. In this case, it would be challenging if there were a need to support new languages. Naively fine-tuning a previously learned TTS model on new languages may not be ideal. The challenges come from several aspects in multilingual TTS scenario: different languages are diverse in linguistic contents and pronunciations [3]. Also, Multilingual corpora usually contain unique speakers for each language [6]. It is thus difficult for the TTS model to preserve distinct pronunciations, language prosodies and speaker identities in previous languages while learning a new language. As a result, synthesis performance on previous languages degrades severely. From the context of lifelong learning, this is known as "catastrophic forgetting" [7].

Alternatively, one can either retrain the TTS model from scratch using both new language data and the original data, or develop co-training strategies to fine-tune the original model [8]. Both approaches require access to new language data and full previous data that is used to train the original TTS system. Hence, although a decent performance is usually possible, it is very expensive and inefficient in terms of data storage and computation. Further, original training data may not be always available due to privacy concerns.

To address the problems, we for the first time propose a lifelong multilingual TTS approach, where each language is treated as an individual task and learned one-at-a-time. We approach the catastrophic forgetting problem via a data replay training scheme, where past samples are cached in a memory buffer that constraints model training on new tasks. We cast the replay process as a supervised learning problem using current language samples and the buffered *small* amount of past languages samples. To address the issue of heavily language-imbalanced training samples, we propose different sampling strategies to take advantages of the full data. A novel dual-sampler framework is employed to combine benefits of both balanced sampling and random sampling. We conducted both objective and subjective evaluations on CSS10 corpus [6]. Results show that the proposed method achieves 43% Mel-Cepstral Distortion (MCD) improvement compared to the fine-tuning baseline, and it also essentially outperforms other lifelong learning baselines.

2. RELATED WORK

Multilingual TTS. In recent years, several multilingual and/or cross-lingual TTS have been proposed. Some aim to establish a TTS system on low-resource languages from rich-resource languages via transfer learning [9, 2]. This results in multiple TTS models for multiple languages. In contrast, another line of works train a *single* TTS model for multilingual purpose [3, 4, 5]. Unlike these methods that train a TTS model on multiple languages jointly, we study a more human-like multilingual TTS learning process via lifelong learning: the TTS model learns multiple languages one-at-a-time. This procedure enjoys the benefits of lower data storage burden and maintenance cost.

Lifelong learning and applications in speech domain. There have been extensive prior works on lifelong learning [10]. These can be grouped into 3 classes: (1) *regularization*-based which consolidates previous knowledge when learning new tasks via an additional regularization term to the loss function [11, 12]; (2) *replay*-based which employs a buffer of old samples for rehearsal/retraining or to constraint the current learning [13, 14, 15, 16]. (3) *parameter*
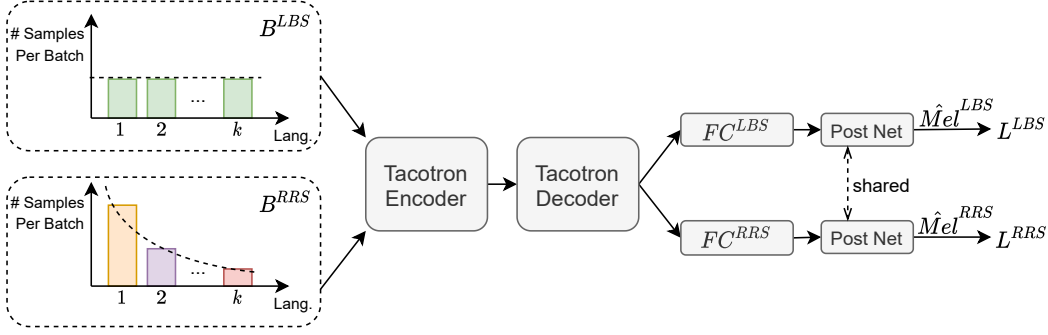
Fig. 1. Proposed imbalance learning method via an auxiliary imbalance sampler. Language embedding, which is concatenated with encoder output is omitted for brevity. During inference, output is obtained from LBS branch.

isolation-based which dedicate different model parameters to each task, and thus dynamically grow the network along with the tasks [17, 18, 19, 20]. In speech domain, a handful of works have applied lifelong learning approaches on ASR [21, 22, 23], multilingual acoustic representation [24], fake audio detection [25] and TTS speaker adaptation [26].

3. METHODS

In this section, we describe the proposed data-replay-based lifelong multilingual TTS approach. We start by introducing the multilingual TTS framework used in this work (Sec. 3.1), followed by describing our formulation of the replay methods (Sec. 3.2).

3.1. Multilingual TTS framework

Our multilingual TTS model is based on Tacotron 2 [27], an attention-based Seq2seq model. We use grapheme as encoder inputs. To account for multilingual control, one-hot encoded language embedding is concatenated with encoder output before fed into the decoder. Decoder output is projected by a fully connected layer and processed by a convolutional post net to generate Log Mel spectrogram (Fig. 1). We follow the network architecture proposed in [3], except that we do not include the residual encoder, adversarial speaker classifier and speaker embedding, since we use a single-speaker corpus and do not intend to perform voice cloning. We use a WaveRNN vocoder [28] pre-trained on the entire CSS10 corpus to convert Mel spectrogram back to waveform.

3.2. Proposed: Multilingual TTS through lifelong learning

Ideally, to easily support new incoming languages, the multilingual TTS learner is expected to preserve a good performance on all languages, when we expose the learner into multiple languages sequentially. For this goal, the continual/lifelong learning algorithms must be able to mitigate the catastrophic forgetting on past languages, without sacrificing performance on the current language. In this study, we consider a data-replay training scheme to address this problem. Formally, let $\mathcal{D}_k$ denote the training set of language k, $\boldsymbol{\theta}_k^*$ denote the optimal model parameters to be learned on language k. We employ a memory buffer $\mathcal{M}_{k-1}$ that stores a limited amount of samples from past languages $1, 2, ..., k - 1$. Suppose that the optimal model parameters $\boldsymbol{\theta}_{k-1}^*$ was learned from past languages $1, 2, ..., k - 1$, our goal is to obtain $\boldsymbol{\theta}_k^*$ from $\mathcal{D}_k$, $\mathcal{M}_{k-1}$ and $\boldsymbol{\theta}_{k-1}^*$:

$$\boldsymbol{\theta}_k^* \leftarrow f(\mathcal{D}_k^+, \boldsymbol{\theta}_{k-1}^*) \qquad (1)$$

where $\mathcal{D}_k^+ = \{\mathcal{D}_k \cup \mathcal{M}_{k-1}\}$, representing the merged training set of $\mathcal{D}_k$ and $\mathcal{M}_{k-1}$. f is a learning process.

Since labeled data is buffered, we explore the feasibility of doing data replay in a supervised learning fashion:

$$L(\boldsymbol{\theta}_k, \mathcal{D}_k^+; \boldsymbol{\theta}_{k-1}^*) = \frac{1}{|\mathcal{D}_k^+|} \sum_{d \in \mathcal{D}_k^+} L(\boldsymbol{\theta}_k, d; \boldsymbol{\theta}_{k-1}^*) \qquad (2)$$

where L denotes the loss function of the TTS model. $\boldsymbol{\theta}_{k-1}^*$ is used to initialize the training of $\boldsymbol{\theta}_k$. This can be formulated as an imbalanced learning problem: training samples from the new language (in $\mathcal{D}_k$) usually significantly outnumber buffered samples from seen languages (in $\mathcal{M}_{k-1}$), making new language the *majority* language, while seen languages *minority/rare* languages. Directly training a TTS model on such highly language-imbalanced samples leads to undesired outcomes - we find that the prosody of the minority languages can be "contaminated" by that from the majority language, making them sound heavily accented (See our audio sample page). Hence, a proper re-sampling strategy for the imbalanced training data is non-trivial. We propose three sampling strategies: random sampling, weighted sampling, and sampling with a dual sampler.

Random Sampling. As the simplest sampling strategy, random sampling does not consider language imbalance. Training samples are *uniformly* sampled. Hence, minority languages in $\mathcal{M}_{k-1}$ receive much less exposure than the majority language in $\mathcal{D}_k$. As a result, random sampling may fail to preserve performance on previous languages.

Weighted Sampling. Instead of uniformly sampling training samples, a straightforward way to handle language imbalance is to sample with language-specific probabilities. Formally, let x_j^i denotes the jth sample of language i, where $i = 1, 2, ..., k$. $C_l = |\{x_j^i | i = l\}|$ denotes the number of samples for language l. Each sample x_j^i in $\mathcal{D}_k^+$ is then assigned with a weight given by the reciprocal of language-wise occurrences: $p_j^i = |\mathcal{D}_k^+|/C_i$, representing the chance that x_j^i is to be sampled. A training batch is formed by sampling from $\mathcal{D}_k^+$ with replacement, using the assigned weights. In this way, rare languages receive higher exposure. Weighted sampling aggressively over-samples rare languages, while under-samples the majority language. This may hinder performance on current language. Also, over-fitting on rare languages may occur.

Dual sampler. On one hand, balancing language-specific samples benefits unbiased representation learning. On the other hand, simply over-sampling may result in over-fitting and affect representation learning of the majority language. Inspired by works on long-tail distribution learning [29, 30], we propose to use a simple yet effec-

Method	Main Training Language																	
	DE	NL				ZH					JA						
	DE	DE	NL	Avg	MCDR	DE	NL	ZH	Avg	MCDR	DE	NL	ZH	JA	Avg	MCDR
Lower bound: Fine-tune	4.11	7.53	4.41	5.97	N/A	8.65	8.11	3.50	6.75	N/A	7.60	7.55	9.66	3.35	7.04	N/A
Upper bound: Joint	3.42	3.42	4.16	3.79	36.52%	3.42	4.16	3.33	3.64	46.12%	3.42	4.16	3.33	3.43	3.59	49.08%
Regularization: EWC	4.11	7.40	4.38	5.89	1.34%	8.22	7.46	3.50	6.39	5.28%	8.18	7.80	8.44	3.48	6.98	0.92%
Replay: GEM	4.11	4.37	4.56	4.47	25.21%	4.33	4.82	4.04	4.27	34.86%	4.71	4.87	4.36	3.68	4.41	37.43%
Rdm. Samp.	4.11	4.39	4.41	4.40	26.30%	4.51	6.04	3.63	4.73	29.98%	4.77	5.09	4.41	3.43	4.43	37.14%
Wtd. Samp.	4.11	4.67	4.95	4.81	19.43%	4.51	4.18	4.38	4.36	35.46%	4.90	4.22	3.57	3.83	4.13	41.34%
Dual Samp.	4.11	4.02	4.30	**4.16**	**30.37%**	4.15	4.40	3.89	**4.15**	**38.52%**	4.56	4.40	3.85	3.25	**4.02**	**42.90%**

Table 1. Test set MCD on seen languages throughout all training stages. Training sequence: DE-NL-ZH-JA. Column Avg denotes the average of MCDs on tested languages. MCD reduction (MCDR) is computed with respect to the lower bound performance (Fine-tune).

tive dual-sampler framework (Fig. 1) to combine benefits from both balanced and imbalanced training samples. To encourage exposure to minority languages, similar to weighted sampling, a language-balanced sampler (LBS), which samples a training batch (B^{LBS}) containing balanced number of samples for each language, serves as the primary sampler to learn unbiased Seq2seq network and projection layer (FC^{LBS}). The regular random sampler (RRS) randomly samples a training batch (B^{RRS}) from the full training samples and is used as an auxiliary sampler to prevent ill-fitting of the majority languages representations. A separate projection layer FC^{RRS} is utilized for RRS task, while the Tacotron Seq2seq network and post net are shared. In this way, the Seq2seq model is exposed to both sampling strategies, thus it takes advantage of full data and mitigates over-fitting on minority languages. The total training loss is computed as:

$$L = \gamma L^{LBS} + \beta L^{RRS} \tag{3}$$

where L^{LBS} and L^{RRS} are Tacotron training loss [27] from LBS and RRS task, respectively. In our experiments, we empirically set $\gamma = 0.5$ and $\beta = 1.0$.

4. EXPERIMENTAL SETUP

Dataset. We use CSS10 [6], a 10-language speech corpus, with a single but different speaker for each language. We select 4 languages to form a task sequence: German (DE), Dutch (NL), Chinese (ZH), Japanese (JA). We follow the released train/validation splits from [4], resulting in 15.1hrs, 11.5hrs, 5.4hrs, 14.0hrs of training audio, respectively. We further split the last 20 samples from each language's validation set as their test sets. We evaluate the lifelong learning approaches using the following language order: DE-NL-ZH-JA. For replay-based methods we use a buffer size of 300 utterances, which roughly correspond to 0.6h audio. Randomly selected samples are pushed to the buffer after training on each language. When new samples need to be buffered, we randomly pop old samples to keep the buffer language-balanced throughout the entire training sequence.

Lower/Upper bounds and baselines. Like prior works [3, 4] do, the upper bound performance can be achieved by jointly training the multilingual TTS model using full data from all 4 languages. This assumes access to the entire multilingual corpus beforehand. In contrast, the fine-tuning baseline uses only current language data to fine-tune a model trained on previous languages. This leads to aforementioned "catastrophic forgetting" and serves as the lower bound.

In addition, we also implemented two other lifelong learning algorithms and compare the proposed methods with them:

- **Elastic Weight Consolidation (EWC)** [12]. EWC is a regularization-based lifelong learning method. It introduces a regularization term to penalize updates on certain model parameters that are important to previously seen languages.

- **Gradient Episodic Memory (GEM)** [15]. Similar to our proposed approach, GEM is also a replay-based method that uses a memory buffer to store past samples. When training on a new language, buffered samples constrain updates via gradient projection, such that losses on past samples do not increase.

Model and hyper-parameters. We train each language for 100 epochs. We adopt Adam optimizer with a initial learning rate of 0.001, decayed by half after 60 epochs. Batch size is set to 84. For all sequential learning methods, the optimal model parameters obtained from the proceeding language is used to initialize current language training.[1]

5. RESULTS

5.1. Objective Evaluation

We use Mel-Cepstral Distortion (MCD) [31] between the synthesized speech and ground-truth speech as the objective evaluation metric. Test set MCDs throughout the training sequence DE-NL-ZH-JA are reported in Table 1. At each training stage, we evaluate MCD on all languages seen so far (e.g. when training on ZH, we evaluate test set MCD on DE, NL and ZH). Averaged MCD on all available test languages at each stage are reported. Following [23], we compute MCD reduction (MCDR) against the lower bound baseline (*Fine-tune*) to measure the recovered performance from the catastrophic forgetting brought by fine-tuning.

First, across all training stages, there are large MCD gaps between the lower bound *Fine-tune* and the upper bound *Joint* (e.g. 7.04 v.s. 3.59 at the final stage), demonstrating the catastrophic forgetting phenomenon. With such large gaps, we found that utterances synthesized by the *Fine-tune* model are incomprehensible. Pronunciations and speaker identity information for past languages are already lost.

---

[1]Audio samples are available at https://mu-y.github.io/speech_samples/llltts/. Code is available at https://github.com/Mu-Y/lll-tts.
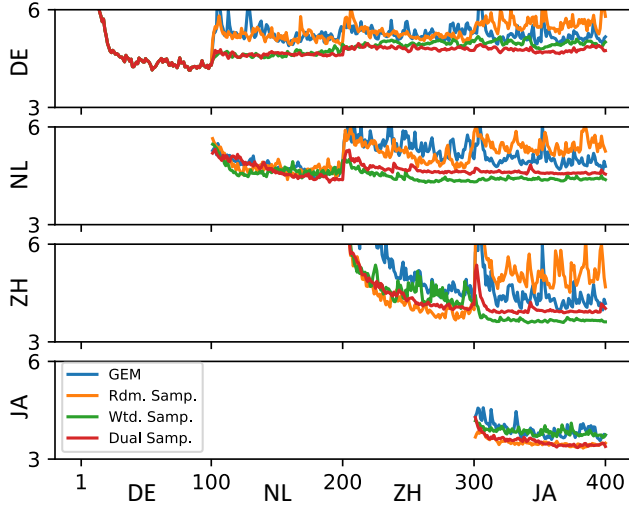
Fig. 2. Dev set MCD for each language throughout the training sequence DE-NL-ZH-JA. Each language is trained for 100 epochs. Curves are plotted with a smooth factor of 0.5.

Second, we can see that all our proposed data replay methods as well as *EWC* and *GEM* outperform the *Fine-tune* baseline, with the largest improvement coming from the proposed *Dual Sampling* approach. Among these methods, *EWC* mitigates forgetting to a very limited extent (with a MCDR up to 5.28% at ZH stage). This is consistent with the findings in [23] which demonstrates that directly imposing constraints on model parameters via regularization may not be optimal. With *Dual Sampling* strategy, the supervised-learning-based replay method outperforms the gradient-based replay method *GEM*. A possible explanation is that ground truth labels of past samples may provide more useful information than gradients, and thus result in superior performance.

Finally, we compare the three proposed sampling strategies. Overall, the proposed *Dual Sampling* achieves the highest MCDR among all replay and regularization methods at all stages. As an example, at the final stage JA, *Dual Sampling* achieves 42.90% MCDR against the *Fine-tune* baseline. Compared to *Random Sampling*, *Dual Sampling* reaches lower MCD on all seen languages so far, indicating the benefit of including a language-balanced sampler into supervised data replay. On the other hand, *Weighted Sampling* is slightly better than *Dual Sampling* on past languages (NL: 4.22 vs 4.40, ZH: 3.57 vs 3.85), at the cost of much higher MCD on the current language (JA: 3.83 vs 3.25). This result is consistent with our expectation: over-sampling minority languages (in this case, DE, NL, ZH in the buffer) may lead to the ill-fitting on the majority language (JA). As a result, *Weighted Sampling* is worse than *Dual Sampling* in terms of average MCD. At JA stage, we can also observe that, despite that *Weighted Sampling* over-samples DE, MCD on DE is in fact higher than *Dual Sampling*. One possible reason is that as the earliest training language, DE receives the most exposure to the TTS model with the aggressive over-sampling strategy imposed by *Weighted Sampling*, making the TTS model over-fit on DE. This result show that the proposed *Dual Sampling* strategy can mitigate over-fitting on minority languages. Similar general patterns can also be observed at other training stages.

5.2. Learning Curve

We show the learning curves (in terms of Dev set MCD) throughout the training sequence DE-NL-ZH-JA in Fig. 2. Due to the inferior

	DE	NL	ZH	JA
GEM	2.52±0.40	2.02±0.26	2.23±0.26	3.69±0.20
Rdm. Samp.	2.11±0.39	1.78±0.36	2.31±0.31	3.64±0.30
Wtd. Samp.	3.23±0.35	**3.13±0.30**	**3.15±0.29**	3.60±0.34
Dual Samp.	**3.28±0.39**	3.01±0.35	3.02±0.28	**3.76±0.35**

Table 2. Test set MOS (with standard deviation) after the final training stage.

performance, we leave out *EWC* and *Fine-tune* and only compare the proposed replay methods plus *GEM*. At DE stage (epoch 1-100), since no past samples are available, all 4 methods are equivalent to a basic supervised learning on DE, and thus 4 curves are identical. At the next stage (NL), DE MCD increases for all methods, with larger margins for *GEM* and *Random Sampling*. This indicates the more severe forgetting phenomenon in *GEM* and *Random Sampling*. We also observe that *Weighted Sampling* landed a slightly higher MCD than *Dual Sampling* on current language (NL). This margin becomes larger as the training move forward (ZH at epoch 200-300, JA at epoch 300-400), demonstrating the negative effect of over-sampling past languages. On the other hand, *Weighted Sampling* produces lower MCD on past languages than *Dual Sampling*, due to its more intensive past languages exposure. Corroborating with results shown in Table 1, we conclude that *Weighted Sampling* is better capable of preserving performance on past languages but is relatively weaker on learning new languages, while *Dual Sampling* yields a more balanced overall performance, as shown by the average MCD scores.

5.3. Subjective Evaluation

We conduct subjective evaluation after the entire training sequence DE-NL-ZH-JA finished. Test set utterances of all 4 languages are synthesized by the final model. We recruit participants on Amazon Mechanical Turk to score the naturalness of the synthesized audio in a scale of 1 (unnatural) to 5 (natural). Each language was evaluated by 15 native speakers. The Mean Opinion Scores (MOS) are shown in Table 2. *Fine-tune* and other methods are left out because they clearly **fail** to synthesize intelligible speech. Consistent with our findings in Table 1 and Fig. 2, *Weighted Sampling* produces more natural speech on past languages (NL, ZH), while sacrificing the performance on the final language (JA). Without a proper sampling strategy, *GEM* and *Random Sampling* both make "mumbled" or heavily accented speech, which are often hard to understand. The results demonstrate the effectiveness of our proposed sampling method to mitigating the forgetting phenomenon.

6. CONCLUSION

We have for the first time presented a lifelong learning approach for multilingual TTS. Our proposed supervised-learning formulation and novel sampling strategies significantly improve synthesis quality over multiple lifelong learning baselines. Nevertheless, we acknowledge that there exist more accents and/or mis-pronunciations in the synthesized utterances compared to the joint training approach. This indicates that this challenging human-like continual language learning problem for a TTS model is far from solved. In future work, we plan to investigate techniques to enhance synthesis naturalness, for example, by dedicating language-specific model parameters. Cross-lingual transfer learning may also be explored for the TTS model to adapt to new languages.

7. REFERENCES

[1] Quanjie Yu, Peng Liu, Zhiyong Wu, Shiyin K Ang, Helen Meng, and Lianhong Cai, "Learning cross-lingual information with multilingual blstm for speech synthesis of low-resource languages," in *ICASSP*, 2016, pp. 5545–5549.

[2] Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," in *Proc. Interspeech 2019*, 2019, pp. 2075–2079.

[3] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *arXiv preprint arXiv:1907.04448*, 2019.

[4] Tomáš Nekvinda and Ondřej Dušek, "One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech," in *Proc. Interspeech 2020*, 2020, pp. 2972–2976.

[5] Jingzhou Yang and Lei He, "Towards universal text-to-speech.," in *INTERSPEECH*, 2020, pp. 3171–3175.

[6] Kyubyong Park and Thomas Mulc, "CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages," in *Proc. Interspeech 2019*, 2019, pp. 1566–1570.

[7] Michael McCloskey and Neal J Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, pp. 109–165. Elsevier, 1989.

[8] Mutian He, Jingzhou Yang, Lei He, and Frank K. Soong, "Multilingual byte2speech models for scalable low-resource speech synthesis," *arXiv preprint arXiv:2103.03541*, 2021.

[9] Younggun Lee, Suwon Shon, and Taesu Kim, "Learning pronunciation from a foreign language in speech synthesis networks," *arXiv preprint arXiv:1811.09364*, 2018.

[10] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[11] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[12] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[13] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[14] Eden Belouadah and Adrian Popescu, "Il2m: Class incremental learning with dual memory," in *ICCV*, 2019, pp. 583–592.

[15] David Lopez-Paz and Marc'Aurelio Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, pp. 6467–6476, 2017.

[16] Gobinda Saha, Isha Garg, and Kaushik Roy, "Gradient projection memory for continual learning," in *International Conference on Learning Representations*, 2021.

[17] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[18] Ju Xu and Zhanxing Zhu, "Reinforced continual learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 907–916.

[19] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3366–3375.

[20] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang, "Long live the lottery: The existence of winning tickets in lifelong learning," in *International Conference on Learning Representations*, 2021.

[21] Samik Sadhu and Hynek Hermansky, "Continual Learning in Automatic Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 1246–1250.

[22] Brady Houston and Katrin Kirchhoff, "Continual learning for multi-dialect acoustic models.," in *INTERSPEECH*, 2020, pp. 576–580.

[23] Heng-Jui Chang, Hung yi Lee, and Lin shan Lee, "Towards Lifelong Learning of End-to-End ASR," in *Proc. Interspeech 2021*, 2021, pp. 2551–2555.

[24] Samuel Kessler, Bethan Thomas, and Salah Karout, "Continual-wav2vec2: an application of continual learning for self-supervised automatic speech recognition," *arXiv preprint arXiv:2107.13530*, 2021.

[25] Haoxin Ma, Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, and Chenglong Wang, "Continual learning for fake audio detection," *arXiv preprint arXiv:2104.07286*, 2021.

[26] Hamed Hemati and Damian Borth, "Continual speaker adaptation for text-to-speech synthesis," *arXiv preprint arXiv:2103.14512*, 2021.

[27] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[28] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[29] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019, pp. 2537–2546.

[30] Junjie Zhang, Lingqiao Liu, Peng Wang, and Chunhua Shen, "To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions," *arXiv preprint arXiv:1912.04486*, 2019.

[31] Robert Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. IEEE, 1993, vol. 1, pp. 125–128.