



Improving Multilingual Text-to-Speech with Mixture-of-Language-Experts and Accent Disentanglement

Jing Wu¹, Ting Chen², Minchuan Chen², Wei Hu², Shaojun Wang², Jing Xiao²

¹Fudan University, China

²Ping An Technology, China

rebecca.wu@foxmail.com, tingc95@outlook.com, cmc_1023@163.com

Abstract

Code-switching and accent control is particularly valuable in multilingual text-to-speech (TTS) systems as both of them contribute to improving the authenticity and comprehensibility. However, the issues of seamless integration of languages within a single utterance and the thorough disentanglement of different attributes without bilingual data remains to be solved. To conquer these problems, a computation-efficient model is proposed in this paper. Firstly, the Mixture of Language Experts (MoLE) module is introduced as the encoder to extract language-specific features and fuse intra-utterance semantic information. Secondly, embedding methods together with several regularization strategies and speaker consistency constraints are utilized to ensure that the generated speech aligns with the desired accent. Experiments show that the proposed model can improve the performance of code-switching accent-controllable multilingual TTS over the baseline model in terms of fluency and naturalness.

Index Terms: text-to-speech synthesis, mixture of experts, multilingual, accent

1. Introduction

Multilingual speech is the bridge of cross-cultural communication. The internationalization of today's world has brought about two common phenomena in languages: code-switching and accent speech. Code-switching refers to the alternating use of multiple languages or their variants within a single conversation or sentence. Accent speech is related to the distinct speech patterns and pronunciation characteristics such as intonation retained by non-native speakers when speaking a second language. Therefore, a practical multilingual TTS system is called for providing natural personalized synthesized voice and minimizing accent and pronunciation problems.

In recent years, TTS technology has undergone significant advancements. There has been remarkable progress in generating high-quality and high-fidelity speech output. Some examples include the well-known Tacotron [1], Tacotron2 [2], FastSpeech [3], FastSpeech2 [4], VITS [5], VITS2 [6], etc. These achievements are primarily driven by advancements in deep learning, speech and natural language processing. In terms of generating personalized and customized voices, efforts have been made on modeling speaker identities [7], styles [8], emotions [9], and accents [10]. One notable area of development is multilingual support. When training a multilingual TTS system, bilingual or multilingual data from a single speaker is scarce. The lack of such data poses great challenges in modeling and decoupling the speaker's identity, language and accent. In order to support code-switching and enhance the controllability of languages and accents, some additional modules are required,

which will increase the computational complexity of the model.

The Mixture of Experts (MoE) [11, 12] structure has achieved great success in deep learning by substantially scaling up the model capacity and introducing only a small amount of computational overhead. The MoE layer consists of a group of experts and a gating network, where experts handle different subsets of inputs, and the gating network determines which inputs are sent to which experts. Most recently, many works [13, 14, 15] on automatic speech recognition (ASR) have employed MoE architectures to be powerful tools to recognize multilingual and intra-sentence code-switching speech.

In this work, a novel TTS model based on MoLE is proposed to enhance the multilingual and code-switching capabilities with less computational resource. Specifically, an end-to-end speech synthesis model, VITS2, is leveraged as the backbone of the proposed model. Based on the VITS2 structure, the Stable and Transferable MoE (ST-Moe) [16] blocks are introduced to replace the feed-forward networks (FFN) in Transformer [17] architecture. In the hope that each expert is language-specific, there are totally $N + 1$ experts within a MoLE block, where N is the number of languages that the TTS model supports. An additional expert is designed to be language-agnostic to support generic modeling. In order to have explicit control over speaker identities, languages and accents, the proposed model is designed delicately to model the accent variance. Three innovative aspects of our work include: (1) We introduce the MoLE module, enabling the synthesis of authentic pronunciation even in multilingual contexts. (2) Our model not only generates authentic pronunciation but also accented speech. (3) We verify that the Speaker Consistency Loss (SCL) and the normalization tricks can enhance speaker similarity and accent controllability.

The rest of the paper is structured as follows: an brief overview of the related techniques is given in Section 2, the detailed architecture of the proposed model is introduced in Section 3, the effectiveness of the proposed model is conformed in Section 4, and conclusion is presented in Section 5.

2. Related Works

2.1. Multilingual TTS

In practical use, multilingual TTS systems need to handle two different types of input scripts: one is that there is only one language within a single utterance, the other is the code-switching case. A simple way to achieve natural multilingual TTS requires collecting speech data from bilingual or multilingual speakers. One example is that a Mandarin-English corpus recorded by a bilingual speaker [18], and then use it to train a bilingual TTS system. Since mixed-lingual corpora are scarce, another way is to leverage monolingual speech data from speakers of different

languages. Considering that every language has its own alphabet, some TTS systems use different symbol sets for each language [19] while others present phonemes with shared International Phonetic Alphabet (IPA) [20]. In terms of discussing the drawbacks, the former approach limits the code-switching consistency and cross-language transfer while the latter approach faces challenges when handling language-specific nuances and phonetic features. With the development of deep learning, an increased number of end-to-end TTS systems have realized the function of multilingual and code-switching speech synthesis by introducing language IDs and an language embedding [21].

2.2. Accented TTS

Controllable accented TTS systems have significant practical value spanning across various applications. For example, they help express specific pronunciation features and offer diverse character voices, enhancing the overall immersive experience for the users. In terms of represent an accent, Variational Auto-encoder (VAE) has been widely utilized. Multi-level VAE [22], conditional VAE (CVAE) [10] and hierarchical CVAE [19] have been proposed to build a control mechanism that disentangle accented pronunciation information from acoustic features. Besides, accent variance adaptor [23] and adaptive weight factorization method [24] are also proved to be effective on multi-accent data sets in a low-resource condition.

3. Method

Motivated by the success of MoE applied to multilingual ASR and the state-of-the-art performance non-autoregressive TTS models achieved, a novel TTS model which takes advantages from these two architectures is proposed in this paper. The overall architecture of the proposed model is shown in Figure 1.

3.1. MoE based Backbone

The basic architecture is VITS2, which makes several improvements based on the VITS model. It mainly consists of below components: a posterior encoder, prior encoder, decoder, discriminator and duration predictor. Through the combination of CVAE, normalizing flows and adversarial training, the model achieves remarkably high-quality speech output at low latency.

Sparkled by the fact that the process of different languages can be achieved by different experts each specializes in a specific domain, MoE is incorporated into the proposed model. Instead of homogeneous experts in most MoE structures [11, 12, 25, 26], the MoLE, which is inspired by [13], is adopted to the proposed model. A MoLE module comprises two types of experts: language-specific experts and a language-agnostic expert. In this paper, two encoders equivalent to an English expert and a Mandarin expert are designed to individually provide language-specific information. Under this condition, the information from different languages is isolated from each other. So a language-agnostic expert is added to strengthen the interaction of code-switching context information. The MoLEs are used in the shared parts, such as the attention layers in the stacked Transformer encoder blocks in both the text encoder and the flow block. In contrast to the standard Transformer blocks, MoLE FFN modules as shown in Figure 2 are introduced to replace the FFN.

In practice, ST-MoE [16] serves as the basis to realize the MoLE architecture on our own. The Router z-loss introduced in ST-MoE, as defined in Equation 1, resolves instability issues in training phase and has a slight improvement in the quality-

stability trade-off issues.

$$L_z(x) = \frac{1}{T} \sum_{i=1}^T \left(\log \sum_{j=1}^N e^{x_j^{(i)}} \right)^2 \quad (1)$$

where T is the number of tokens, N is the number of experts, and $x \in R^{T \times N}$ are the logits going into the router. Secondly, rather than the balancing loss, a language identity (LID) loss inspired by [14] is used to train the gating network. The token-wise probabilities output by each gating network p_{lid} are used to calculate the cross-entropy (CE) loss jointly with the ground truth LID sequences u_{lid} . The detailed depiction is stated as:

$$L_{lid} = CE(p_{lid}, u_{lid}) \quad (2)$$

According to the top-1 choice, data samples are routed to the corresponding language-specific expert and the common language-agnostic expert. Thus, the computational complexity will not increase regardless of the number of experts as only two experts will be routed at each time.

3.2. Accent Control

Speaker encoder is a learnable lookup table that encodes the speaker identity into a speaker code. Similarly, the accent encoder is also a learnable lookup table which encodes an accent category into an accent embedding. Since the speaker identity and accent are highly correlated, it is of great importance to promote disentanglement between these two embeddings. Drawing upon the demonstrated robustness, the feature decoupling idea introduced in VICReg [27] is incorporated in our model. A loss function with variance-covariance regularization terms is used in our proposed model.

To give a brief interpretation, let $E \in R^{n \times d}$ be an embedding table, where n represents the number of embedding vectors and d denotes the dimensionality of each embedding vector. Then the variance regularization term, formulated as Equation 3, is defined as a hinge function on the standard deviation of E along dimension d to ensure that each component $e^j \in R^{n \times 1}$ is differentiated. The covariance regularization term, defined as Equation 4, is designated to drive the off-diagonal elements of the covariance matrix $C(E)$ to approach zero, which requires that the components in one embedding vector should not have redundant information about each other.

$$L_{var} = \frac{1}{d} \sum_{j=1}^d \max \left(0, \gamma - \sqrt{\text{var}(e^j)} + \epsilon \right) \quad (3)$$

$$L_{covar} = \frac{1}{d} \sum_{i \neq j} [C(E)]_{i,j}^2 \quad (4)$$

where $\text{var}(\cdot)$ is the variance metric, γ represents a fixed target value for the standard deviation and ϵ is set to safeguard against numerical instabilities.

Besides, the cross-correlation criterion [20] is proposed to decorrelate random variables in the speaker embedding table E^S and accent embedding table E^A , which is defined by:

$$R^{AS} = \frac{1}{B-1} \left(\tilde{E}^A - \mu_{E^A} \right) \left(\tilde{E}^S - \mu_{E^S} \right)^T \quad (5)$$

$$L_{xcorr} = \frac{1}{D_a D_s} \sum_{i,j} \left(R_{i,j}^{AS} \right)^2$$

where R^{AS} is the cross-correlation matrix, $\tilde{E}^A$ and $\tilde{E}^S$ are sampled from accent and speaker identity embedding tables within a batch of size B respectively, μ_{E^A} and μ_{E^S} are the mean vectors computed from the embedding tables.

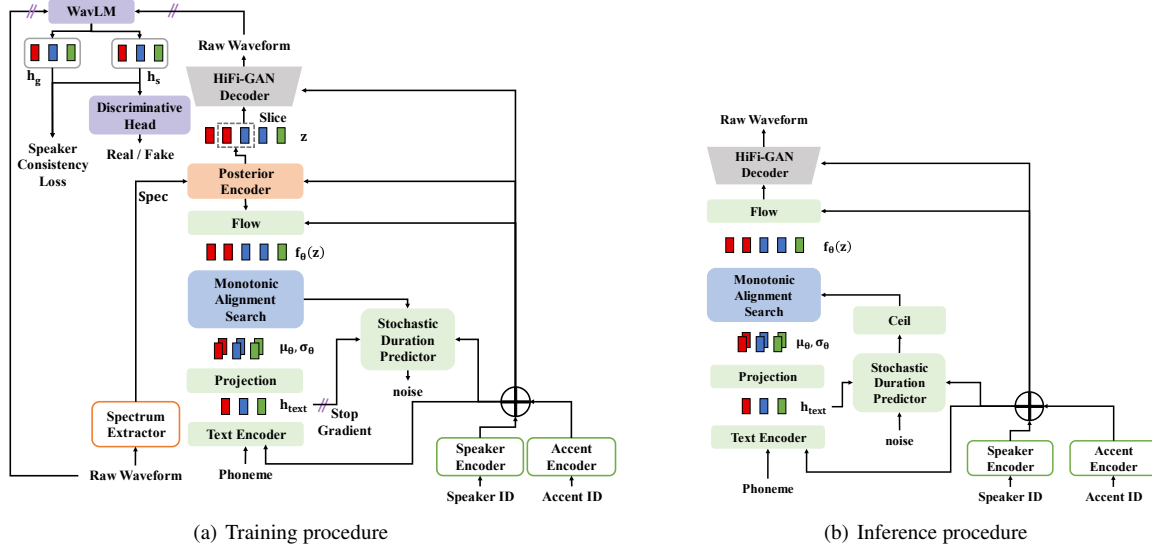


Figure 1: System diagram of the proposed model. (a) Training procedure and (b) Inference procedure.

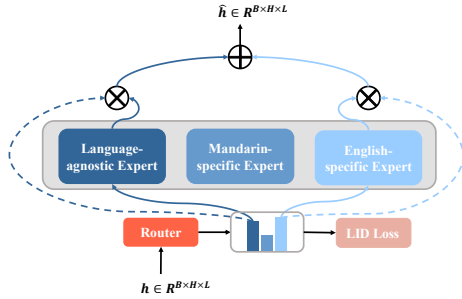


Figure 2: MoLE with two language-specific experts and one language-agnostic expert. The module takes hidden variable h as input, and outputs $\hat{h}$.

3.3. Speaker Consistency Loss

In order to alleviate the inconsistency of speaker timbre in synthesized speech caused by incomplete decoupling properties, we follow YourTTS [28] to add the SCL in the final loss. A pretrained speaker encoder model is used to extract the speaker-related embedding from both the ground truth audio g and the synthesized audio s . The SCL is designed to maximize the cosine similarity between these two embeddings and the equation is as follows:

$$L_{scl} = 1 - \frac{\alpha}{n} \sum_{i=1}^n \cos_sim(\phi(g_i), \phi(s_i)) \quad (6)$$

where α is a positive weight, $\cos_sim$ is the cosine similarity-based retrieval, ϕ is the function extracting the speaker embedding from a given audio. Besides, the newly introduced StyleTTS2 [8] uniquely transfers knowledge from speech language models (SLMs) to speech generation tasks via SLM discriminators with the intention of enhancing the naturalness of synthesized audio.

Therefore, our proposed model is built by combing the SCL with SLM discriminators to achieve both spoken content modeling and speaker identity preservation. Specifically speaking, the pretrained WavLM [29] is employed as the discriminators. And apart from the discriminative head, an additional speaker verification head was appended to compute the SCL.

4. Experiments

4.1. Datasets

Experiments were conducted on two languages - English and Mandarin. For English, VCTK [30] dataset was utilized. For Mandarin, AISHELL-3 [31] dataset was used. All audios were resampled to 44100 Hz and preprocessing techniques were applied to ensure consistent sample loudness and to eliminate extended periods of silence. For each speaker, utterances were partitioned into training, validation, and test set at a ratio of 8 : 1 : 1. Besides, to evaluate the cross-lingual performance with only monolingual training data, a commercially available bilingual female voice dataset from DataBaker¹ is supplemented. Since the speaker is native in Mandarin, only the Mandarin data is used to fine-tune the model and the code-switching data was selected for code-switching evaluation. In each code-switching utterance, the words in the base sentence were randomly replaced with words from another language.

4.2. Training setup

Following [6], We trained our networks using the AdamW [32] optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and weight decay $\lambda = 0.01$. The learning rate decay was scheduled by a 0.999 factor in every epoch, with an initial learning rate of 2×10^{-4} . All models were trained using mixed precision training on four NVIDIA V100 GPUs with a batch size of 32.

4.3. Experimental results

In this paper, both subjective and objective metrics are used to obtain a more comprehensive evaluation. Character Error Rate (CER) and Speaker Encoder Cosine Similarity (SECS) are chosen to give a objective evaluation, while Mean Opinion Score (MOS) and ABX preference are relied on to evaluate naturalness and accent-controllability. For CER, the publicly accessible multilingual Whisper Large-v3 version² is utilized to transcribe our synthesized speech and then the transcription results are compared with the original text. For SECS, the embeddings of both the synthesized audio samples and reference ones are

¹<https://www.data-baker.com/>

²<https://github.com/openai/whisper>



Figure 3: *ABX preference test of accent-controlled speech. "A" refers to the synthesized speech with a native accent, while "B" represents the opposite.*

first computed by the pretrained speaker encoder H/ASP model [33], and then the resulting embeddings are used for cosine similarity-based retrieval. We used the vanilla VITS2 model as the baseline. Besides, twelve participants who speak fluently both in English and Mandarin are invited to conduct ABX and MOS tests (each model underwent tests with 20 samples). Audio samples can be found online³.

4.3.1. Evaluation in monolingual cases

In order to evaluate how well the model performs in monolingual cases, ablation studies were conducted on the testing set from AISHELL-3 and VCTK. In this case, only one language is contained in an utterance. Table 1 demonstrates results via objective measurement.

Table 1: *Objective results in monolingual cases. Accent Control module is abbreviated as AC. The proposed model can be denoted as: (B)+MoLE+SCL+AC.*

Model	CER(↓)		SECS(↑)	
	English	Mandarin	English	Mandarin
GT	1.24	8.34	/	/
Baseline (B)	3.58	18.17	59.08	69.03
(B) + AC	2.43	18.02	56.98	67.97
(B) + SCL	4.00	19.10	65.93	73.78
(B) + AC + SCL	2.75	16.85	63.83	71.78
(B) + MoLE	3.13	17.14	63.05	72.23
(B) + MoLE + SCL	3.59	15.65	62.85	72.77
Proposed model	2.92	15.41	66.22	74.71

From Table 1, it can be seen that when empowering the baseline VITS2 model with accent controllability, SECS is decreased in both languages. This is mainly caused by the disentanglement of personal properties. Since embedding vectors are specific representations of speaker identities and accents, they may be unable to fully capture individual differences in respect to voice characteristics and pronunciation habits. Besides, because only monolingual data is available in the training phases, two embeddings are not adequately trained and there is a subtle overlap between them. To make up for this problem, SCL is therefore added. It is obvious that the introduction of SCL can eliminate the problem of reduced speaker similarity caused by accent control and makes an improvement in the speaker timbre preservation when compared with the baseline. In addition, the result of CER indicates that MoLE is capable of enhancing performance by first separately processing language-specific information and then integrating it together.

Besides, the result of the proposed model's ABX preference test is displayed in Figure 3. It can be seen that the participants mainly preferred the speech synthesized with a native accent as they sounded more natural and fluent, thus proving that the accent controllability is an improvement to multilingual TTS.

Furthermore, a t-SNE projection is shown in Figure 4. Comparing Figure 4(b) and Figure 4(c), it can be observed that the introduction of regularization terms for accent-control improves the disentanglement of accent states. Besides, the clustering result of speaker identity becomes more explicit when SCL is adopted as displayed in 4(a).

³<https://eagerr1102.github.io/demo/>

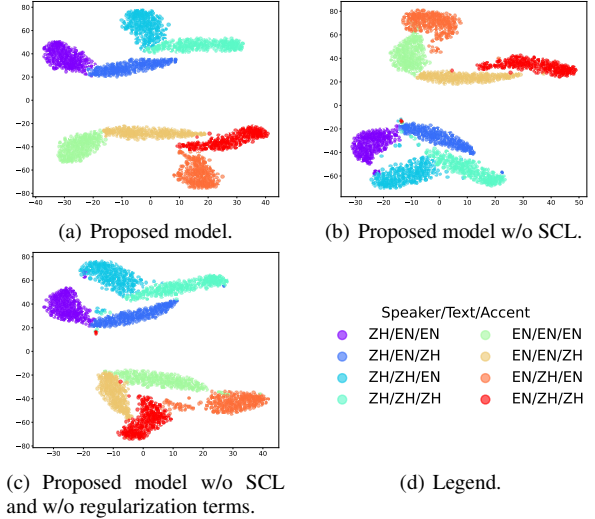


Figure 4: *t-SNE visualization of the effect of voice cloning and accent control, using 2D PCA of embeddings. The legend is shown in Figure 4(d) where ZH represents Mandarin and EN stands for English.*

4.3.2. Evaluation in code-switching cases

In this section, ablation studies are made to find out the effects of MoLE and SCL on synthesizing code-switching utterances. Results are shown in Table 2. Compared CER achieved by models with and without MoLE, there is no doubt that the incorporation of MoLE effectively improves the model's ability to handle the language-specific information. Besides, when SCL is added to the model, both CER, SECS, and MOS are decreased. It might be caused by the speaker encoder. Whats more, when the model is equipped with accent controllability, both CER and MOS are improved, which indicates that accent control can help to improve the overall naturalness and the intelligibility of the code-switching synthesized speech. Furthermore, by comparing Table 1 and Table 2, CER is worse while SECS is better when conducted on code-switching cases. This might be because the model was fine-tuned on a single-speaker dataset for code-switching cases, thus resulting in better speaker timbre retention compared to a multi-speaker dataset.

Table 2: *Results in code-switching cases.*

Model	CER(↓)	SECS(↑)	MOS(CI)
GT	8.58	/	4.33(± 0.06)
Baseline (B)	37.85	79.98	3.60(± 0.08)
(B) + MoLE	27.20	82.44	3.89(± 0.09)
(B) + MoLE + SCL	25.58	82.09	3.86(± 0.08)
Proposed model	25.48	82.06	3.99(± 0.07)

5. Conclusion

In this paper, a MoLE-based model is proposed to enhance the performance of multilingual TTS system in monolingual and cross-lingual cases. By combining the router z-loss and language-routing mechanism, the performance of cross-lingual TTS is improved compared to the baseline. The proposed model also outperforms the baseline in explicitly disentangling the speaker identity, language, and accent. Controllable speech output can be achieved while preserving the target speaker's voice in all desired languages and accents. Besides, since MoLE consists of a gate to choose the best language-specific expert, the model offers efficient inference with no extra overhead when supporting more languages. In the future, more code-switching-specific approaches and accent-control-specific methods will be explored to boost the overall performance of multilingual TTS system.

6. References

- [1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.
- [2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.
- [4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.
- [5] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
- [6] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design," in *Proc. Interspeech*, 2023, pp. 4374–4378.
- [7] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," in *Proc. Interspeech*, 2019, pp. 2105–2109.
- [8] 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.
- [9] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *Proc. ICASSP*, 2021, pp. 5734–5738.
- [10] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "Accented text-to-speech synthesis with a conditional variational autoencoder," *arXiv preprint arXiv:2211.03316*, 2022.
- [11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [12] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.
- [13] Y. Kwon and S.-W. Chung, "Mole: Mixture of language experts for multi-lingual automatic speech recognition," in *Proc. ICASSP*, 2023, pp. 1–5.
- [14] W. Wang, G. Ma, Y. Li, and B. Du, "Language-Routing Mixture of Experts for Multilingual and Code-Switching Speech Recognition," in *Proc. Interspeech*, 2023, pp. 1389–1393.
- [15] F. Tan, C. Feng, T. Wei, S. Gong, J. Leng, W. Chu, J. Ma, S. Wang, and J. Xiao, "Improving End-to-End Modeling For Mandarin-English Code-Switching Using Lightweight Switch-Routing Mixture-of-Experts," in *Proc. Interspeech*, 2023, pp. 4224–4228.
- [16] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus, "St-moe: Designing stable and transferable sparse expert models," *arXiv preprint arXiv:2202.08906*, 2022.
- [17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [18] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (mandarin-english) tts," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009.
- [19] L. Ma, Y. Zhang, X. Zhu, Y. Lei, Z. Ning, P. Zhu, and L. Xie, "Accent-vits: accent transfer for end-to-end tts," in *National Conference on Man-Machine Speech Communication*. Springer, 2023, pp. 203–214.
- [20] R. Badlani, R. Valle, K. J. Shih, J. F. Santos, S. Gururani, and B. Catanzaro, "Radmm: Multilingual multiaccented multi-speaker text to speech," in *Proc. Interspeech*, 2023, pp. 626–630.
- [21] S. Li, B. Ouyang, L. Li, and Q. Hong, "Light-tts: Lightweight multi-speaker multi-lingual text-to-speech," in *Proc. ICASSP*, 2021, pp. 8383–8387.
- [22] J. Melechovsky, A. Mehrish, D. Herremans, and B. Sisman, "Learning accent representation with multi-level vae towards controllable speech synthesis," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 928–935.
- [23] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable accented text-to-speech synthesis," *arXiv preprint arXiv:2209.10804*, 2022.
- [24] T.-N. Nguyen, N.-Q. Pham, and A. Waibel, "Syntacc: Synthesizing multi-accent speech by weight factorization," in *Proc. ICASSP*, 2023, pp. 1–5.
- [25] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," in *International Conference on Learning Representations*, 2021.
- [26] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 5232–5270, 2022.
- [27] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *International Conference on Learning Representations*, 2022.
- [28] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.
- [29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [30] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.
- [31] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech*, 2021, pp. 2756–2760.
- [32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.
- [33] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.