



# FaVC: A Validated, Transcribed, Parallel Farsi Speech Dataset for Voice Conversion

Mina Serajian<sup>1</sup>, Saeed Najafzadeh Rahaghi<sup>1</sup>, Hadi Veisi<sup>1</sup>, Saman Haratizadeh<sup>1</sup>

<sup>1</sup> School of Intelligent Systems, University of Tehran, Iran  
Mina.serajian@ut.ac.ir, najafzadeh.sd@ut.ac.ir, h.veisi@ut.ac.ir,  
haratizadeh@ut.ac.ir

## Abstract

In this paper, we created the first transcribed, parallel, balanced Farsi dataset (FaVC) that can be used for all tasks of speech synthesis, including both parallel and non-parallel voice conversion. FaVC is a balanced dataset that provides all phonemes of the Farsi language in all possible phonetic combinations. A metadata file is provided, including Farsi transcriptions, normal text, and IPA phoneme transcriptions of all audio files. The first Farsi voice conversion results using FaVC are also reported in this paper. The results indicate that by using FaVC, performance is as good as the chosen baseline methods for voice conversion in English. Moreover, objective evaluations of a voice conversion system that requires a parallel speech corpus can be performed using this dataset.

**Index Terms:** Voice Conversion, Farsi Speech Dataset, FaVC, Speech Synthesis

## 1. Introduction

### 1.1. History

Voice Conversion (VC) is a branch of speech synthesis that aims to convert speech characteristics in different aspects. It has achieved advances such as changes in voice identity, emotion [15], accent [16], and cross-language voice conversion [3]. VC applications include dubbing movies with the main speaker's voice, personalization in text-to-speech (TTS) systems [16], detecting Automatic Speaker Verification (ASV) attacks, and medical applications such as improving speech intelligibility for dysarthric patients [16].

Based on training data, VC models and datasets can be classified into parallel and non-parallel types. In parallel datasets, similar utterances in the same language with the same content are spoken by several source and target speakers, but in non-parallel datasets, these source and target utterances do not have the same content or even the same length [2]. Additionally, in some datasets, the training and test data are in the same language, but in others, the data are in different languages [3].

Statistical methods for voice conversion primarily involve modeling speech features to align the source voice with the target voice (common techniques include Gaussian Mixture Models). For non-parallel data, techniques like autoregressive (AR) models or probabilistic tensor networks are employed to bridge data gaps [1]. Various methods have been proposed to solve the problems associated with statistical methods. The use of deep learning has significantly developed in artificial intelligence during recent years. Modern methods also integrate statistical models with deep learning to enhance the accuracy and quality of voice conversion. The introduction of deep

learning, particularly deep neural networks (DNNs), generative adversarial networks (GANs), and variational autoencoders (VAEs), has improved VC systems' performance. These models allow for more robust feature extraction, better mapping between source and target speakers, and higher-quality speech synthesis [5]. CycleGAN-VC and StarGAN-VC are both deep learning models based on GANs, designed for non-parallel VC, enabling speech transformation without requiring paired datasets. CycleGAN-VC introduces a method for voice conversion without requiring parallel data, leveraging CycleGAN [6].

In many articles, voice conversion methods have been used to improve text-to-speech systems because the goal of both (VC and TTS systems) is to produce high-quality speech that aligns with appropriate linguistic content [12]. Researchers have tried to use these methods to convert voice. Applying deep learning in VC results in very high accuracy compared to previous methods, but the main problem in deep learning for voice conversion still remains, which is the lack of proper and sufficient data [1].

### 1.2. Data-related Issues in Deep Learning

Deep learning models are data-hungry [19]. Without proper data, even the best models cannot perform well. Four common issues related to data in deep learning are [18][22]:

- **Data quantity:** Deep learning models typically require a large-enough amount of [labeled] data to learn effectively. Insufficient data can lead to poor generalization and overfitting. There is no high-quality speech dataset in Farsi for speech processing tasks. Additionally, when data are imbalanced, the model might become biased toward the majority class, resulting in poor performance on the minority class. The ideal data for training a speech processing system should have an equal number of male and female speakers. Furthermore, in the application of cross-lingual VC, there is a need for Farsi data in addition to other languages.
- **Data quality:** Low-quality data, such as noisy, incomplete, or inconsistent data, can degrade model performance. The model may learn incorrect patterns, leading to unreliable predictions. There is no reliable and available collection of noise-free Farsi phonemes with all related IPA phoneme transcriptions. Additionally, sometimes data distribution in training differs from real-world scenarios, reducing the generalization ability. In speech, it is highly desirable to have a dataset where the distribution of phonemes is the same as in daily conversations and that covers all phonetic combinations, which is currently lacking in Farsi. Moreover, using a regular microphone makes the audio recordings in the

Table 1: A selection of Speech datasets (~ indicates the approximate number)

Feature\Dataset	VCC 2018 [4]	VCC 2020 [3]	VCTK [28]	LibriSpeech [16]	DeepMine [28]	Farsi ESD [24]	Farsdat [25]	FaVC
# paired utterances	81	20 in English	0	0	0	0	0	405
# non paired utterances	81	50 English 70 non English	~ 43600	~ 287367	36713	470	More than 2000	0
# speakers	8	8 English 8 non English	109	2484	67	2	300	11
Parallel / non parallel	Semi parallel	Semi parallel	Non parallel	Non parallel	Non parallel	Non parallel	Non parallel	Parallel
Language	En	En, Fi, Ch, Du	En	En	Fa	Fa	Fa	Fa
Duration (sec)	1 to 6	1 to 6	2 to 6	2 to 16	more than 30	10 to 30	2 to 20	1 to 8
SR (kHz)	22.05	24	48	16	22.05	22.05	44	22.05
Main Task	VC	VC	TTS	ASR/ TTS	TTS	Emotion	ASR	VC

dataset more applicable to real-world scenarios. Besides that, ethical and privacy concerns, especially regarding sensitive data, pose legal and societal challenges.

### 1.3. Why FaVC?

Using VC objective evaluations such as MCD [27], PCC [28], and RMSE [29] on F0, FFE [30], and GPE [31] requires ground truth, which means that in VC, the reference speaker must utter the same content as the source speaker. These evaluations are only possible when a parallel dataset exists.

The availability of speech datasets for the Farsi language is quite limited, with only a few datasets currently available. However, these datasets are not designed specifically for speech conversion purposes, creating a considerable gap in resources for Farsi-language research. Public datasets such as Common Voice [8] still aren't parallel. Additionally, each of these datasets suffers from data-related issues, as mentioned in Section 1.2.

To address these challenges, we decided to create our own Farsi speech database (FaVC), which is balanced, open-source, noiseless, sufficiently large for VC, and transcribed. To evaluate and analyze the quality of our dataset, five voice conversion models were trained with FaVC. Besides voice conversion, due to the similar structure of FaVC to standard datasets such as VCTK [28], it can be used in other speech processing tasks such as few-shot TTS or ASR research.

## 2. Related work

### 2.1. Non-Farsi Datasets

Numerous datasets have been used in voice conversion research, including the VCC2018 dataset, designed for non-parallel voice conversion with recordings from 12 speakers, each providing 80 English utterances [4]. The VCC2020 dataset includes recordings of 70 English sentences and some in German, Finnish, and Chinese for cross-lingual conversion [3]. The CSTR VCTK Corpus offers 44 hours of data from 109 speakers with diverse English accents [9], while LibriSpeech provides 1,000 hours of phonetically diverse audiobook

recordings from over 2,000 speakers, with precise text alignments [11].

Single-speaker datasets like LJSpeech, though high-quality, are unsuitable for voice conversion due to the need for multi-speaker training [21] in VC. Table 1 summarizes the key characteristics of these datasets.

### 2.2. Farsi Datasets

Currently, no publicly available datasets are specifically prepared for converting speech in Farsi, despite the Persian language being spoken by over 100 million people worldwide and ranking among the top 20 most widely spoken first languages. The DeepMine dataset [23] addresses this gap by providing 120 hours of high-quality, multi-speaker TTS data from 67 speakers, sourced from audiobooks on the IranSeda website [28]. This dataset has enabled the training of multi-speaker TTS synthesizers and a vocoder, achieving high naturalness scores [28].

Similarly, the Farsi ESD dataset [19] focuses on emotional speech recognition, containing audio samples representing 5 emotions. This dataset is carefully annotated and validated to support the development of accurate emotional speech recognition systems.

The Large FarsDAT dataset [20] offers over 2,400 hours of transcribed Farsi speech data, representing diverse speakers, accents, and contexts to support Automatic Speech Recognition (ASR). However, its use in TTS applications requires additional phonetic transcription and alignment processes.

The Common Voice corpus [8], an open-source multilingual dataset under a CC-0 license, provides transcribed speech for ASR and other domains, although it is non-parallel and includes varying speaker contributions in different recording environments. The available background noise in the dataset makes it suitable for real-world applications, but a significant problem remains: the lack of validated, accurate, noise-free, phonemically transcribed paired data that can be used for multiple purposes, such as modeling the correct way of speaking all phonemes and more. Table 1 summarizes the aforementioned datasets.

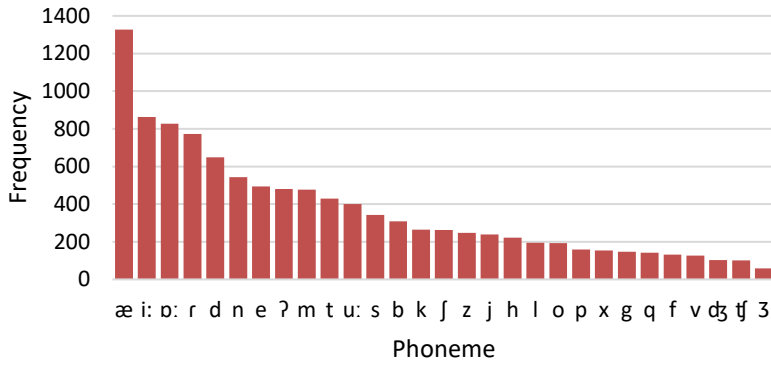


Figure 1: The histogram of Farsi phoneme (IPA format) counts in the FaVC dataset.

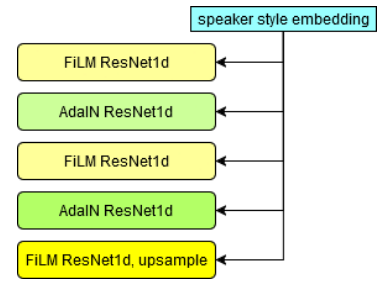


Figure 2: FiLM and AdaIN on Decoder of TTS.

### 3. FaVC Details

In this section, parts of FaVC and its features are described.

#### 3.1. FaVC Data

As FarsDat [20] brings balanced sentences, they were selected to be uttered for our voice dataset. All sentences were made to consist of all Farsi phonemes to allow dialect comparison among speakers [20]. To record the voices, a microphone with medium characteristics (making the voices closer to reality) was used, with an SNR of at least 36 dB. The distance between the speakers and the microphone was between 12 and 16 cm. Therefore, no special equipment or laboratory environment was used for sound recording. All the sounds were recorded in a home environment without special noise. Five women and six men (one is different and labeled as an outlier) with different age groups were selected. These people had no accent, and Farsi was their mother tongue. Their voices were continuously recorded with Adobe Audition 2022 software in the same environmental conditions, and then the file of each sentence was extracted in this software and saved in a folder with the name of the speaker number. The name of each file was saved as "speaker number\_sentence number." The sampling rate is 22.05 kHz. A metadata file is prepared for FaVC, which includes the number of sentences, Farsi text transcription, translation of the sentences into English, and IPA phoneme transcription of each sentence. FaVC is about 220, which is larger than the VCC 2018 [4] dataset and makes it suitable for speech processing tasks.

#### 3.2. Statistics of the Dataset

There are 10,656 phonemes in 405 sentences of the FarsDAT dataset. In this section, some statistics from the voice dataset of this research are shown, which may be useful.

##### 3.2.1. Duration of Voice Files

Figure 3 shows the histogram of the duration of the audios. There are 4,455 audios uttered by 11 speakers, each saying 405 paired sentences. The audios range from 1 to 8 seconds. Most of the files have a duration between 1 and 6 seconds, containing a complete sentence, which is totally aligned with the existing benchmark datasets like VCTK [28] and makes FaVC suitable

for almost all hardware resources. There are only 10 audios with a length of more than 7 seconds.

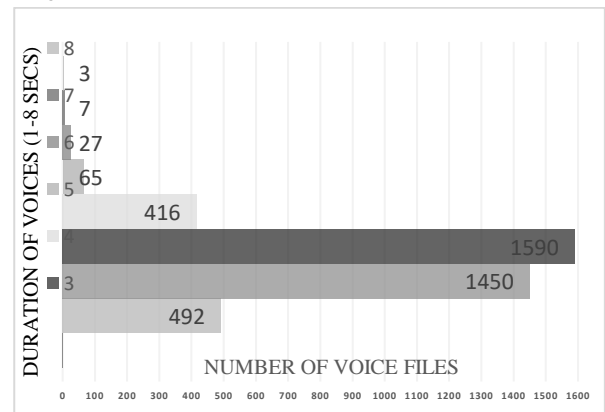


Figure 3: The duration distribution of the recorded audios.

##### 3.2.2. Phoneme Coverage and Being Balance

Figure 1 depicts the histogram of Farsi phonemes' counts in all sentences of the FarsDAT dataset. FarsDAT is phonetically balanced, with sentences covering all phonetic combinations in Farsi [25].

#### 3.3. Speaker Details

The number of men and women is quite equal. Farsi was their mother tongue and they had no dialect. There are 6 men and 5 women who uttered 405 identical sentences. The ages of the females are 19, 21, 37, 40, and 56. The ages of the men are 21, 23, 33, 35, 39, and 58.

## 4. Experiments and Results

To assess the quality of our dataset and its suitability for voice conversion tasks, we trained five VC models using the FaVC.

#### 4.1. Architecture of Models

The first model was CycleGAN-VC [6], which was a one-to-one model. Other models were based on StarGAN's idea. The three models were StarGAN-VC [7], StarGAN-VC2 [10], and StarGANv2-VC [9]. CycleGAN-VC leverages the primary challenge in voice conversion [1], which is mapping features between source and target speakers, particularly in cases where paired training data is unavailable. This problem is solved by adopting a cycle-consistency loss. Unlike methods that require separate models for each source-target pair, StarGAN-VC enables scalable and flexible VC by utilizing a single unified framework. StarGANv2-VC significantly improves upon StarGAN-VC by introducing a disentangled representation of speaker identity and style, enabling diverse, natural-sounding voice conversions.

Another approach for VC is leveraging TTS. StyleTTS-VC [26] is an advanced one-shot VC framework that utilizes a novel method of learning disentangled representations through transfer learning from StyleTTS [25]. A modified StyleTTS-VC is proposed (FiLM+AdaIN with modified cycle-consistency loss), capable of converting speech from any source speaker to an arbitrary target speaker using only a brief reference audio clip from the target speaker. A combination of FiLM [32] and AdaIN [33] is proposed in the TTS decoder architecture (the first stage of training, and the architecture is the same as in [26]); then, for more data augmentation in the VC phase (the second stage of training [26]), a modified cycle-consistency loss is proposed. We added the second term to the loss function as shown in Equation (1).

$$\mathcal{L}_{\text{cycle}} = \mathbb{E}_{x,t} \left[ \left\| x - G(h, S(x), p_{\text{trg}}, n_{\text{trg}}) \right\|_1 + \left\| \hat{x} - G(h, S(x), p_{\text{trg}}, n_{\text{trg}}) \right\|_1 \right] \quad \text{Equation(1)}$$

All the notations are the same in [26].

#### 4.2. Training

Eight speakers were included in training with 350 fixed sentences, and the remaining 55 sentences were placed in the validation set, while the rest of the speakers were placed in the test set. To perform an objective evaluation (MCD), all the sentences of all speakers were used. The configuration of all implementations is the same as the officially released code of the selected models.

All aforementioned models output mel-spectrograms. To convert them into a voice signal, we leverage HiFi-GAN as our vocoder. To avoid using all speakers in vocalization, we fine-tuned it with FaVC in a non-parallel way by selecting five random speakers who uttered 81 different and distinct FaVC sentences.

#### 4.3. Evaluation

To evaluate the converted voices, Mel-Cepstral Distortion (MCD) and two types of Mean Opinion Score (MOS), MOS-P (similarity between speakers) and MOS-N (assessing naturalness), were used as objective and subjective evaluation methods, respectively [1][23]. We asked 20 native Farsi speakers to rate the results on a scale of 1 to 5. For MOS-N, 1 means "Bad" and 5 means "Excellent," and for MOS-P, 1 means "Not similar" and 5 means "The same speaker."

As some papers did not report the selected evaluation methods, and their training datasets differed, making comparison impossible, we trained all models twice with the original paper's configuration settings. First, they were trained

with an English dataset, then the models were trained with FaVC. Afterward, we conducted an evaluation using the selected metrics. All results are reported in Table 3.

Table 2: Evaluations of Models

Model	Dataset	MCD	MOS-P	MOS-N
<b>Modified StyleTTS-VC</b>	VCTK	-	3.66	3.89
	FaVC	5.58	3.87	3.94
<b>StarGAN-VC</b>	VCC2018	7.1	2	3.5
	FaVC	7.66	2.9	3.4
<b>StarGAN-VC2</b>	VCC2018	7	1.89	3.3
	FaVC	7.32	2.6	3.3
<b>CycleGAN-VC</b>	VCC2018	7	2.1	3.1
	FaVC	7.2	2.5	3.3
<b>StarGANv2-VC</b>	VCC2018	5.39	3.8	3.9
	FaVC	6.05	4.1	3.8

## 5. Discussion

All models trained on the FaVC data have successfully achieved high-quality results in MOS-N, MOS-P and MCD criteria, comparable to those obtained with English datasets. This demonstrates that FaVC can be placed as a strong benchmark dataset for voice conversion.

Moreover, the effect of the modified cycle consistency loss function in the training of the StyleTTS-VC encoder and the combination of FiLM and AdaIN in the StyleTTS decoder for one-shot learning is demonstrated in Figure 4.

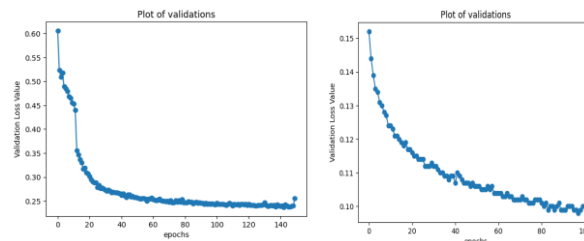


Figure 4: Modified StyleTTS-VC Loss Function Training - Left: Stage 1 (TTS Decoder) and Right: Stage 2 (Mel Encoder).

## 6. Conclusion

In this paper, we described the voice recording process and statistics for the FaVC dataset. This dataset covers most data-related issues of deep learning and is comparable to mostly English datasets, which are commonly used in VC research papers. To assess the quality of this dataset and establish a baseline for future research, we utilized four deep learning VC architectures based on GAN and TTS. The evaluation results indicate that the quality of the converted voices was comparable to the original paper findings.

For future work, we recommend producing a multi-language dataset by recording English translations of FaVC sentences. Additionally, labeling will be added, as it is necessary for ASR and TTS research. The FaVC dataset for voice conversion and speech synthesis models in Farsi is freely available upon request for research purposes.

## 7. References

- [1] B. Sisman, J. Yamagishi, S. King and H. Li, "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132-157, 2021, doi: 10.1109/TASLP.2020.3038524
- [2] Mingyang Zhang, Berrak Sisman, Li Zhao, Haizhou Li, "DeepConversion: Voice conversion with limited parallel training data", *Speech Communication, Elsevier*, vol. 122, 2020, pp. 31-43, ISSN 0167-6393.
- [3] Zhao, Y., Huang, W. C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., ... & Toda, T., "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion", Submitted to ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020.
- [4] Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T.H., & Ling, Z., "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods", *Odyssey*, 2018.
- [5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," in *IEEE Trans. Audio, Speech Lang. Process.*, IEEE, vol. 15, no. 8, Nov 2007, pp. 2222–2235.
- [6] Kaneko, Takuhiro, and Hirokazu Kameoka. "CycleGAN-vc: Non-parallel voice conversion using cycle-consistent adversarial networks." 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018.
- [7] Kameoka, Hirokazu, et al. "StarGAN-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks." 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.
- [8] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., & Weber, G. (2019). Common Voice: A Massively-Multilingual Speech Corpus. *ArXiv, abs/1912.06670*.
- [9] Li, Yinghao Aaron et al. "StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion." *Interspeech* (2021)
- [10] Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N., "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion", in *Proc. Interspeech*, 2019, pp. 679-683.
- [11] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.
- [12] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and Wavenet," Mar 2019, arXiv:1903.12389
- [13] M. Zhang, Y. Zhou, L. Zhao and H. Li, "Transfer Learning from Speech Synthesis to Voice Conversion With Non-Parallel Training Data," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, ACM/IEEE vol. 29, 2021, pp. 1290-1302.
- [14] Tyagi, A. K., & Rekha, G. (2020). "Challenges of Applying Deep Learning in Real-World Applications". In *Challenges and Applications for Implementing Machine Learning in Computer Vision* (pp. 92-118). IGI Global.
- [15] Kun Zhou, et al, "Emotional voice conversion: Theory, databases and ESD", *Speech Communication, Elsevier*, February 2022, Volume 137, Pages 1-18, ISSN 0167-6393.
- [16] Y. Zhao, M. Kuruvilla-Dugdale and M. Song, "Voice Conversion for Persons with Amyotrophic Lateral Sclerosis," in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2942-2949, Oct. 2020, doi: 10.1109/JBHI.2019.2961844.
- [17] Whang, S.E., Roh, Y., Song, H. et al. "Data collection and quality challenges in deep learning: a data-centric AI perspective". *The VLDB Journal* 32, 791–813 (2023). <https://doi.org/10.1007/s00778-022-00775-9>
- [18] Kun Zhou, Berrak Sisman, Mingyang Zhang, Haizhou Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion", in *INTERSPEECH 2020*, October 25–29, 2020, pp.3416-3420.
- [19] Niloofar Keshiari, Michael Kuhlmann, Moharram Eslami and Gisela Klann-Delius, "Recognizing emotional speech in Farsi: A validated database of Farsi emotional speech (Farsi ESD)", *Springer, Behavior Research Methods* 47, 2015, pp.275–294.
- [20] Bijankhan, Mahmood et al. "FARSDAT- THE SPEECH DATABASE OF FARSI SPOKEN LANGUAGE." (1994).
- [21] keithito/LJ-Speech-Dataset/
- [22] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis". In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1428, 17022–17033.
- [23] Adibian, Majid and Zeinali, Hossein and Barmaki, Soroush, "Deepmine-Multi-Tts: A Farsi Speech Corpus for Multi-Speaker Text-to-Speech". Available at SSRN: <https://ssrn.com/abstract=4673655> or <http://dx.doi.org/10.2139/ssrn.4673655>
- [24] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion", in *IEEE Trans. Speech Audio Process*, IEEE, vol. 6, no. 2, Mar 1998, pp. 131–142.
- [25] Li, Y. A., Han, C., & Mesgarani, N. (2025). StyleTts: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*.
- [26] Li, Y. A., Han, C., & Mesgarani, N. (2023, January). StyleTts-vc: One-shot voice conversion by knowledge transfer from style-based tts models. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 920-927). IEEE.
- [27] R.Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Commun., Comput. Signal Process.*, vol. 1, pp. 125–128, 1993.
- [28] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction Speech Processing*, Berlin, Germany: Springer-Verlag, 2009, pp. 1–4.
- [29] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.
- [30] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 3969–3972.
- [31] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Commun.*, vol. 50, no. 3, pp. 203–214, 2008.
- [32] Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018, April). Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
- [33] Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 1501-1510).