Exploiting voice conversion in creating new TTS voices

Tijana Nosek[1] , Siniša Suzić[1], Nikola Simić[1], Milan Sečujski[1], Darko Pekar[2] and Vlado Delić[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[2] AlfaNum Ltd., Novi Sad, Serbia

tijana.nosek@uns.ac.rs and sinisa.suzic@uns.ac.rs

***Abstract*—The development of high-quality Serbian text-to-speech (TTS) systems for new speakers traditionally requires extensive phonetic and prosodic annotations, a process that is both time-consuming and resource-intensive. This paper explores a novel approach that leverages voice conversion (RVC) techniques to generate synthetic speech in the voice of a target speaker. In this scenario phonetically and prosodically annotated transcriptions of the source speaker could also be re-used for target speaker, i.e. RVC synthesized speech, as well. Four models were evaluated: two adapted with natural speech (30 and 3 minutes, respectively), and two adapted with 2.5h of RVC-generated speech based where RVC conversion models are trained using also 30 and 3 minutes of target speakers' speech. Listening tests assessed speech naturalness and vocal similarity. Results indicate that RVC-generated data enables effective adaptation of multispeaker TTS models, outperforming direct adaptation with very limited natural data. Moreover, the amount of target speaker data used to train the RVC model had minimal impact on final synthesis quality. These findings highlight the potential of using RVC for low-resource speaker adaptation in TTS systems for Serbian.**

***Index Terms*—text-to-speech-synthesis, voice conversion, annotation**

I. Introduction

Text-to-speech (TTS) technologies have rapidly evolved, driving significant advancements in human–machine interaction across various applications. A number of systems achieves the Mean Opinion Score (MOS) comparable to the one achieved for natural speech [1-3]. Many of the models enable the synthesis in the voice of arbitrary speaker using limited amounts of training speech data. This can be achieved by leveraging adaptation steps [4, 5] or even without finetuning, in a technique known in literature as zero-shot TTS [6-9].

However, the majority of recent progress has been centered on high-resource languages such as English and models requiring large amounts of training data, counting in hundreds or even thousands of hours of speech [7], for creating base models. For this reason under-resourced languages, such as Serbian, continue to pose unique challenges for TTS development.

Our research group has already developed a robust TTS engine for Serbian [10], which will be described in more details in the following sections. This system relies on a combination of expert blocks and some recent achievements in TTS based on deep neural networks, and requires phonetically and prosodically annotated text for training. Since the development of annotated datasets is both a time-consuming and an expensive process, this remains a key obstacle in the creation of new TTS voices. There have been attempts in creating of end-to-end systems for Serbian [11], which do not require prosodic annotation of training data and thus decrease data preparation time, but the results are still inferior to [10], since the amount of quality training data is limited.

To address the limitation in creating new voices for Serbian TTS, we propose employing a voice conversion approach to synthesize training data. By leveraging an already annotated speaker as the source, our method converts this voice to simulate new target voices. This innovative strategy circumvents the labor-intensive process of annotating new data from scratch. In doing so, it provides an effective means of expanding the voice repertoire for Serbian TTS without the typical resource constraints.

The remainder of the paper is organized as follows. Section 2 reviews the literature on voice conversion and TTS speaker cloning, highlighting the challenges in low-resource environments. Section 3 describes our proposed methodology in detail. Section 4 presents experimental evaluations and results, and Section 5 concludes the paper with a discussion on implications and future directions for research.

II. Methods

A. DNN based TTS

The highest-quality and most widely used Serbian TTS consists of three primary modules. The first is the front-end, which normalizes the text, performs phonetization (which is a comparatively easy task in Serbian, due to its phonemic orthography), creates prosodic tags for normalized text and then generates sets of linguistic features for each phoneme. Linguistic features represent a set of answers to yes-no questions regarding phonemic and prosodic information for every phoneme (i.e. "Is this phoneme A?", "Is this phoneme accented?", etc). The second module is a DNN-based block that uses linguistic features to predict acoustic features for each phoneme, and the final module is the vocoder. An initial version of the system, which uses the deterministic WORLD vocoder, is described in

[12], while an updated version, which uses the neural HiFi-GAN vocoder is detailed in [10]. Although an universal HiFi-GAN model could be used in the pipeline, we have shown that finetuning the universal model on the data for specific speaker improves overall results.

Within the DNN-based component for acoustic feature prediction, there are two neural networks, as outlined in [12]. One network is dedicated to predicting the durations of phonemes, and the other predicts output acoustic features using both the linguistic inputs and the information about phoneme duration. Linguistic inputs are the same for the both networks. Each of these networks consists of three feed-forward layers with RELU activation, followed by one LSTM layer, and supports multi-speaker modelling. The base multi-speaker models, duration and acoustic, enable better and faster creation of models for new target speaker by finetuning rather training everything from scratch [13]. In the training phase, linguistic features are extracted from phonetically and prosodically annotated databases, while the front-end is used only in the inference phase.

B. Voice conversion with RVC

Voice conversion (VC) is a technique for modification of vocal characteristics of a source speaker's voice so as to closely resemble those of a target speaker, while ensuring the original linguistic content remains unchanged. Unlike TTS systems that generate speech from text, VC algorithms commonly take an existing audio input (the source voice) and convert it to sound like a different speaker (the target voice) while preserving the original speaker's intonation, prosody, and emotional nuances.

Generating new speech content for the purpose of data augmentation within this research is based on an open source project named Retrieval-Based-Voice-Conversion (RVC) [14]. It offers intuitive interface (WebUI) for pragmatic voice conversion to the wider public. RVC is a non-parallel voice conversion system, i.e. it does not require aligned or paired data between source and target speakers. Instead, it only uses speech data from the target speaker during training to build a voice conversion model. This model can then be used to convert speech from an arbitrary source speaker into the voice of the target speaker. The RVC project integrates state-of-the-art architectures like HuBERT or WavLM for feature extraction and employs retrieval mechanisms to map source audio to target speaker embeddings. HuBERT is a transformer-based model trained to predict masked Mel Frequency Cepstral Coefficients (MFCCs) from audio [15].

RVC has been used as a data augmentation tool recently, to create personalized datasets and improve ASR model performance in low-resource language scenarios, particularly for Hindi [16–17]. Here we utilize RVC to generate speech content for data augmentation phase, in order to produce additional training data in new voices. We selected CREPE [18] as the pitch extraction algorithm, ensuring better quality at the expense of computational complexity. Median filter is not applied in the next step as we did not use 'harvest' algorithm for pitch extraction algorithm. We have set the search feature ratio at 0.5 and we proceeded without resampling in the post-processing phase as source and target speaker files had the same sampling rate, matching the options provided by RVC. The volume envelope scaling was set to 1, whereas 0.33 was chosen for protecting voiceless consonants and breath sounds as a default option. Although it is recommended to use at least 10 minutes of low-noise speech data during the training phase, we conducted experiments using 3 minutes and 30 minutes of training data, aiming to explore and compare the performance limits of the RVC algorithm in the case of a tiny target speaker dataset versus a relatively large one.

III. OUR APPROACH

The main limitation of the state-of-the-art Serbian TTS system, as presented in Section II.A, is the requirement for phonetic and prosodic annotation of speech material for a new speaker. This process is extremely time-consuming – often requiring several dozen times the duration of the original speech – and demands expert knowledge, making the creation of new voices expensive and labor-intensive.

In contrast, numerous models with zero-shot speaker adaptation have emerged in recent years and have demonstrated strong performance [6-9]. However, these models are not designed for the Serbian language. In fact, to the best of our knowledge, there are currently no models that can be easily adapted to Serbian, while existing end-to-end TTS models that do not require annotated datasets still exhibit performance inferior to the system described in Section II.A [10]. Therefore, in this work, we investigate the use of RVC to artificially generate sufficient speech material with the voice of the target speaker. In RVC conversion the source speaker is a speaker for which sufficient amount of annotated data already exists. The converted speech data is then used to adapt the multispeaker TTS model presented in Section II.A.

A. Database

Two datasets were used in this study. The first, referred to as the source speaker, contains 2.5 hours of speech (excluding silences) from Speaker A, phonetically and prosodically annotated. The second dataset, referred to as the target speaker, consists of 30 minutes of speech (excluding silences) from Speaker B, with both phonetic and prosodic annotations. Both speakers are female, and all recordings were conducted in a professional studio environment. The audio recordings are sampled at 22.05 kHz with a 16-bit depth.

As the baseline for the adaptation approaches, multispeaker DNN-based TTS model (MS TTS) trained on a total of approximately 20 hours of speech data from 11 different speakers (including the source speaker but excluding the target speaker) was used. It is described in detail in Section II.A.

B. Models

In further experiments, four different models will be evaluated. Theydiffer in data used for training.

The model that will be referred to as *30min_reg* in the remainder of the paper is based on adaptation of the multispeaker TTS (MS TTS) model using 30 minutes of, phonetically and prosodically annotated speech from the target speaker. The adaptation follows a standard fine-tuning

procedure involving both the duration and acoustic models, with the pre-trained parameters of the MS TTS model serving as initialization. The adaptation is performed in two stages: in the first stage, only the speaker embedding is trained while the rest of the model remains frozen; in the second stage, all model parameters are fine-tuned. This two-stage approach has proven effective for building new TTS voices [19], although its performance heavily depends on the availability of annotated material from the target speaker.

The model, hereafter referred to as *3min_reg*, follows the same adaptation procedure but uses only 3 minutes of annotated speech from the target speaker. At the cost of a certain decrease in synthesis quality, the primary advantage of this approach lies in the significantly reduced annotation effort, provided an expert annotator is available.

The third and fourth models, hereafter referred to as *30min_rvc* and *3min_rvc* respectively, also involve adaptation of the MS TTS model, but in these cases, for adaptation the synthetic speech generated via RVC models is used. Two RVC models were trained using unannotated audio recordings of the target speaker: one using the full 30-minute dataset, and another using only 3 minutes. These RVC models were then used to convert the 2.5-hour annotated dataset of the source speaker into synthetic speech in the voice of the target speaker. Since the RVC conversion preserves the linguistic content and prosody of the original speech while transferring the vocal characteristics to match the target speaker, the original annotations were assumed to remain valid. As a result, we obtained 2.5 hours of phonetically and prosodically annotated data in the target speaker's voice, suitable for adapting the MS TTS model. Synthetic speech (2.5h) produced by the RVC model trained on 30 minutes of target speaker audio is used for training *30min_rvc* model, while synthetic speech (2.5h) produced by the other RVC model, the one trained on 3 minutes of target speaker audio, is used for training *3min_rvc* model.

It should be noted that this approach raises questions regarding the quality of the converted speech and the fidelity with which the target speaker's characteristics are preserved. It also puts focus on the requirement that the impact of the amount of target speaker data used to train the RVC model on the final TTS quality should be evaluated.

The DNN-based TTS model used in this study requires a vocoder to generate waveform outputs from the synthesized acoustic features. For all four experiments, we employed the universal HiFi-GAN (HFG) vocoder [20]. Although it is possible to adapt the HFG model to the target speaker, even in the case of Serbian, given sufficient natural speech data [10], we opted not to do so for any of the models. This decision was made to ensure that the evaluation results reflect the quality of the TTS models themselves, without additional influence from vocoder adaptation. It should be noted that the 30 minutes of natural speech from the target speaker should be enough for effectively adapting HFG model, so as 2.5 hours of RVC-generated speech. Such adaptations have the ability to reduce vocoder-induced artifacts in the synthesized speech. Nevertheless, 3 minutes of target speaker data is insufficient for effective HFG adaptation, which further justifies our decision to use the universal HFG model uniformly across all four models and listening tests.

IV. Evaluation

To evaluate the quality of the speech synthesized by the four models presented in the previous section, four listening tests were conducted. As the models *30min_rvc* and *3min_rvc* are not expected to preserve the prosody of the target speaker, objective evaluation metrics were deemed inappropriate for this study. The first two listening tests are designed to assess the overall quality of the synthesized speech, while the third test focuses on evaluating how well the vocal characteristics of the target speaker are preserved in the synthesized output. All listeners were native Serbian speakers.

A. MOS test

The aim of the first listening test was to evaluate the naturalness of the synthesized speech using a 5-point Mean Opinion Score (MOS) scale. A total of 20 listeners participated in the test, and each was asked to rate 15 utterances on a scale from 1 to 5, where a score of 5 indicated highly natural, human-like speech, and a score of 1 indicated completely unnatural or unintelligible speech. Among the 15 samples, 3 were recordings of natural speech from the target speaker, included as reference samples for the highest quality, although this was not disclosed to the listeners. The remaining 12 samples were synthesized using the four models described in Section III.B, with 3 utterances generated per model. While some sentences were shared across models (as regards content), no two models used an identical set of three sentences.

The results are presented in Fig. 1. As expected, natural speech samples achieved the highest average score 4.85. The *30min_reg* model achieved the highest MOS among the synthetic outputs (3.92), which is also expected. The *3min_reg* model received a lower average score (3.12), which is consistent with our previous findings on the limitations of adapting with minimal annotated data. The models based on RVC-generated speech achieved MOS scores of 3.43 (*30min_rvc*) and 3.28 (*3min_rvc*), suggesting a moderate degradation in naturalness when synthetic speech is used for adaptation.

These results indicate that although using RVC-generated data introduces some quality degradation compared to the adaptation with natural speech, it still outperforms adaptation with a small amount of natural speech. Interestingly, the

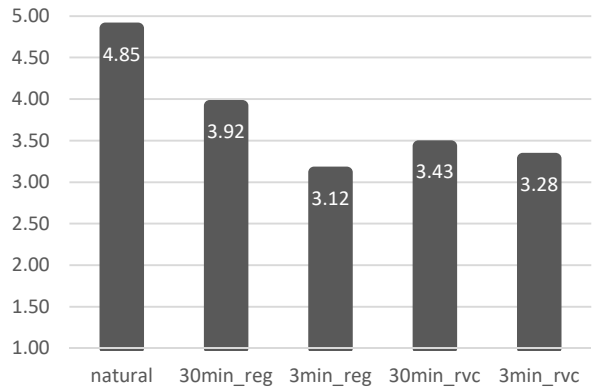

Fig. 1. Results of MOS test – quality of synthesized speech.

relatively small difference between *30min_rvc* and *3min_rvc* (0.15) suggests that the amount of target speaker data used to train the RVC model may not be a critical factor for the quality of converted synthetic speech used for adaptation of MS TTS.

B. Preference test

The second listening test was a preference test designed to provide further insight into the perceived quality of synthesized speech, specifically in terms of listener preference between different synthesizers. A total of 20 listeners participated in the test, and each was presented with 12 pairs of utterances. For each pair, the utterances were identical in lexical content but synthesized using two different models. The listeners were asked to choose the utterance that sounded better in terms of intelligibility and naturalness. A "No preference" option was also provided for cases in which the listener could not decide.

Each of the following model pairings was tested using three pairs of utterances: *30min_rvc* vs. *30min_reg*, *3min_rvc* vs. *3min_reg*, *30min_reg* vs. *3min_reg*, and *30min_rvc* vs. *3min_rvc*.

As shown in Figure 2, the results align well with those of the first listening test. The *30min_reg* model was clearly preferred over both *30min_rvc* and *3min_reg*, reaffirming its superior capacity for high-quality synthesis. Additionally, the *3min_rvc* model was preferred over *3min_reg*, indicating the effectiveness of the RVC-based adaptation approach when only a small amount of target speaker data is available. Notably, there was no clear preference between *30min_rvc* and *3min_rvc*, which suggests that the amount of target speaker data used to train the RVC model does not significantly impact the quality of voice conversion for this task.

These findings support the conclusion that, while the quantity of annotated target speaker data is critical for effective adaptation of the MS TTS model, it is less critical for training the RVC model. Moreover, the results highlight the potential of the proposed approach in low-resource scenarios: the *3min_rvc* model was preferred over *3min_reg* in 60% of the cases, with no preference expressed in an additional 10%, demonstrating its clear advantage when only a minimal quantity of data is available.

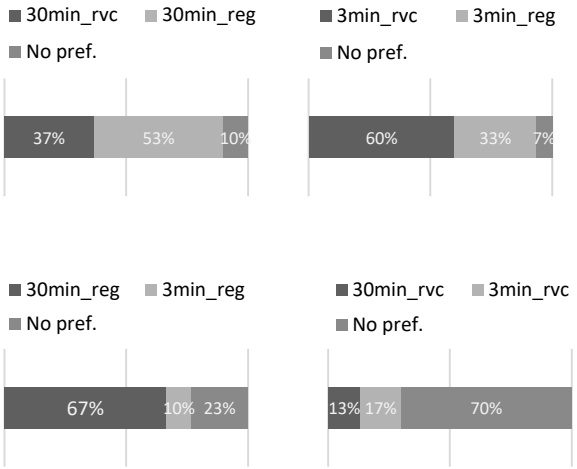

Fig. 2. Results of the preference test – quality of synthesized speech.

C. Similarity test

To evaluate how well the synthesized speech preserves the vocal characteristics of the target speaker, a third listening test was conducted. The goal of this test was to assess the perceived similarity between synthesized utterances and natural recordings of both the target and source speakers. Ideally, synthesized speech should closely resemble the target speaker while minimizing the resemblance to the source speaker—particularly in the case of the *30min_rvc* and *3min_rvc* models, where such interference is more likely due to the use of RVC converted data for adapting MS TTS.

A total of 20 listeners participated in this test. Each listener evaluated 16 sentence pairs on a 5-point similarity scale, where a score of 1 indicated that the utterances were clearly spoken by different speakers, and a score of 5 indicated that the listener is sure that the same speaker was perceived in both utterances. In half of the sentence pairs, synthesized utterances were compared to natural recordings of the target speaker, and in the other half, they were compared to natural recordings of the source speaker. Each model contributed 4 synthesized utterances, 2 were evaluated for similarity to the target speaker, and 2 for similarity to the source speaker. For half of the pairs, the reference and synthesized utterances had identical content, while the other half differed in content. It was concluded that the variation in content did not significantly affect listeners' perception of speaker similarity.

The results, presented in Fig. 3, show that the *30min_reg* and *3min_reg* models produce speech that does not resemble the source speaker (both getting mean similarity scores of 1.5), while achieving moderate similarity to the target speaker (3.2 and 3.1, respectively). In contrast, the *30min_rvc* and *3min_rvc* models produce speech that is less similar to the target speaker (2.6 and 2.1, respectively). Interestingly, while *30min_rvc* shows higher similarity to the source speaker (2.2) compared to the other models, both RVC-based models maintain relatively low resemblance to the source speaker overall (2.2 and 1.5 for *30min_rvc* and *3min_rvc*, respectively).

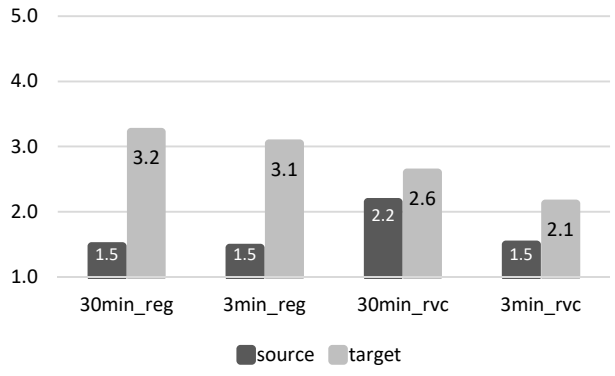

Fig. 3. Results of similarity test – voice similarity between natural and synthesized utterances.

These findings suggest that although the RVC-based models do not adequately capture the target speaker's identity, they also do not strongly retain the source speaker's characteristics, resulting in speech that lies somewhere in between. The reason why *3min_rvc* is rated as less similar to the source speaker than *30min_rvc* is unclear and warrants further investigation through more extensive testing.

V. Conclusion

This study demonstrates that RVC-based voice conversion is a viable strategy for adapting multispeaker TTS models when only a minimal amount of natural speech data from a new target speaker is available. RVC-generated speech enables the re-use of existing phonetically and prosodically annotated transcriptions from other speakers data, eliminating the need for expert manual annotation and significantly reducing development costs. While the highest synthesis quality was still achieved using 30 minutes of manually annotated natural speech, RVC-based models consistently outperformed those adapted using only 3 minutes of such data. Importantly, the quality of RVC-generated data did not degrade substantially when the RVC model was trained on just 3 minutes of target speaker audio. These findings suggest that RVC can play a crucial role in the development of high-quality Serbian TTS voices in low-resource scenarios. Future work will explore methods to improve speaker identity preservation in RVC-generated speech and to integrate a vocoder adapted to the target speaker for further enhancement of synthesis quality.

Acknowledgment

This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7449, Multimodal multilingual human-machine speech communication, AI-SPEAK.

References

[1] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan and R.A. Saurous, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions", IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018, Calgary, AB, Canada, pp. 4779-4783

[2] J. Kim, J. Kong and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech", International Conference on Machine Learning, 2021, pp. 5530-5540

[3] H. Kim, S. Kim and S. Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance", International Conference on Machine Learning, 2022, Baltimore, MD, USA, pp. 11119-11133

[4] S.F. Huang, C.J. Lin, D.R. Liu, Y.C. Chen and H.Y. Lee, "Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech" IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 2022, pp.1558-1571

[5] S. Kim, H. Kim and S. Yoon, "Guided-TTS 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data" arXiv preprint arXiv:2205.15370, 2022

[6] E. Casanova, J. Weber, C.D. Shulby, A.C. Junior, E. Gölge, E. and M.A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone", International conference on machine learning, 2022, , Baltimore, MD, USA, pp. 2709-2720

[7] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li and L. He, "Neural codec language models are zero-shot text to speech synthesizers", arXiv preprint arXiv:2301.02111, 2023

[8] Z. Jiang, J. Liu, Y. Ren, J. He, Z. Ye, S. Ji, Q. Yang, C. Zhang, P. Wei, C. Wang, C. and X. Yin, "Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis", arXiv preprint arXiv:2307.07218, 2023

[9] 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, 36, 2023, pp.19594-19621

[10] S. Suzić, D. Pekar, M. Sečujski, T. Nosek, V. Delić, "HiFi-GAN based Text-to-Speech Synthesis in Serbian", 30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, pp. 2231-2235

[11] T. Nosek, S. Suzić, M. Sečujski, V. Stanojev, D. Pekar and V. Delić, "End-to-end speech synthesis for the Serbian language based on Tacotron", International Conference on Speech and Computer (SPECOM 2024), 2024, Belgrade, Serbia, pp. 219-229

[12] T. Delić, M. Sečujski, S. Suzić, "A review of Serbian parametric speech synthesis based on deep neural networks", Telfor Journal, vol. 9, no. 1, 2017, pp. 32 – 37

[13] T. Delić, S. Suzić, M. Sečujski, D. Pekar,"Rapid Development of New TTS Voices by Neural Network Adaptation", 17th International Symposium INFOTEH-JAHORINA, 2018, Jahorina, Bosnia and Herzegovina, pp. 1-6

[14] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI, accessed: April 2025.

[15] 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

[16] A. Tathe, A. Kamble, S. Kumbharkar, A. Bhandare, and A.C. Mitra "Transcription and translation of videos using fine-tuned XLSR Wav2Vec2 on custom dataset and mBART". arXiv preprint arXiv:2403.00212.

[17] A. Kamble, A. Tathe, S. Kumbharkar, A. Bhandare and A.C. Mitra, "Custom Data Augmentation for low resource ASR using Bark and Retrieval-Based Voice Conversion", arXiv preprint arXiv:2311.14836.

[18] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 161-165,

[19] M. Secujski, D. Pekar, S. Suzic, A. Smirnov and T. Nosek, "Speaker/Style-Dependent Neural Network Speech Synthesis Based on Speaker/Style Embedding", J. Univers. Comput. Sci., 26(4), 2020, pp.434-453.

[20] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis", Advances in neural information processing systems, 33, 2020, pp.17022-17033.