Subtitle Synthesis in the Target Language with Prosody Modification

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Most of the videos, either in the educational or the entertainment sector, are recorded only in one language (native language), and subtitles are provided for the other languages (foreign languages). The foreign audience has to focus on both the video and the subtitles displayed on the top or bottom of the screen. It becomes more complicated when the audience has to simultaneously see the content written by the speaker and the subtitles displayed on the screen (video lectures). In this paper, we propose a subtitle synthesis approach, which takes the subtitles in the target language and uses a text-to-speech synthesis system that can modify the prosody and synthesize the speech in the target language with the desired prosodic characteristics and incorporate it in the source video. The duration of the synthesized speech (audio track) should match with the duration of the video to mix them. The obtained mean opinion scores say that the translated videos are quite natural, and the naturalness can be further improved by lip-syncing the video with the audio. There are several issues to be addressed in the subtitle synthesis approaches which can enhance the quality of experience. This approach has numerous applications in the education and the entertainment sector.**

***Index Terms*—Subtitle synthesis, Prosody modification, Speech-to-Speech translation, Text-to-Speech synthesis, Machine translation**

I. Introduction

Text-to-Speech synthesis refers to the process of synthesizing artificial speech samples corresponding to the given text. Modern speech synthesis systems are capable of synthesizing natural-sounding expressive speech when conditioned on the desired prosody or emotion. We can even modify the prosody by changing the prosodic parameters like fundamental frequency (f_0), duration of the phones, and energy. This makes the speech synthesis systems useful for numerous practical applications like human-machine interactions, screen readers, subtitle synthesis, etc. In this paper, we use the text-to-speech system for the subtitle synthesis approach.

Subtitle synthesis is the task of synthesizing the speech in the target language (referred to as target audio track) from the subtitle text of the source video and incorporating the synthesized target audio track in the source video (recorded in the native language). If the target language subtitles are not available, we use the automatic speech recognition (ASR) system to transcribe the source audio track into the source subtitle text. Then, we use the Machine translation (MT) system to translate the source subtitle text into the target language subtitle text. Subtitle synthesis has numerous applications in the educational and entertainment sectors since most of the videos are recorded only in the native language. Subtitles are provided for the foreign languages instead of re-recording the video in the foreign languages. The foreign audience has to simultaneously watch the video and read the subtitles displayed on the screen, which limits the effectiveness of watching the video. So, we propose subtitle synthesis, an approach that synthesizes speech in the target foreign language from the subtitles and incorporates the synthesized foreign audio track in the source video.

First, the video track and the audio track are separated from the source video. Subtitles of the foreign target language are fed to the text-to-speech synthesis system to generate the target language audio track. The prosody of the synthesized target language audio track should match the prosody of the source audio track to have a natural experience. So, in order to incorporate the source prosody, the text-to-speech synthesis system should have the provision to modify the prosodic parameters like fundamental frequency (f_0), duration of the phones, and the energy. Most of the state-of-the-art text-to-speech synthesis systems like Tacotron [1], TransformerTTS [2], Deep-Voice [3], Char2Wav [4] don't have the provision to modify the fundamental frequency, duration, etc. Further, most of them are the autoregressive based approaches, which have higher inference times during the synthesis. Some of the systems use autoregressive neural vocoders like WaveNet [5], WaveRNN [6], LPCNET [7] to generate speech signals from the melspectrograms. The proposed subtitle synthesis method uses Prosody-TTS system [8], a non-autoregressive text-to-speech synthesis system that has the provision to modify the prosody by changing the prosodic parameters.

Since the prosody of the synthesized speech depends on the parameters like duration of the phones, fundamental frequency (f_0), and energy, they are modified to incorporate the target prosody. The fundamental frequency is modified to ensure that the emotion in the synthesized speech matches the emotion in the source video. Durations of the phones are modified to ensure that the silence regions in the synthesized target language audio track match the source audio track. Phrases of the source and target language audio track should be of

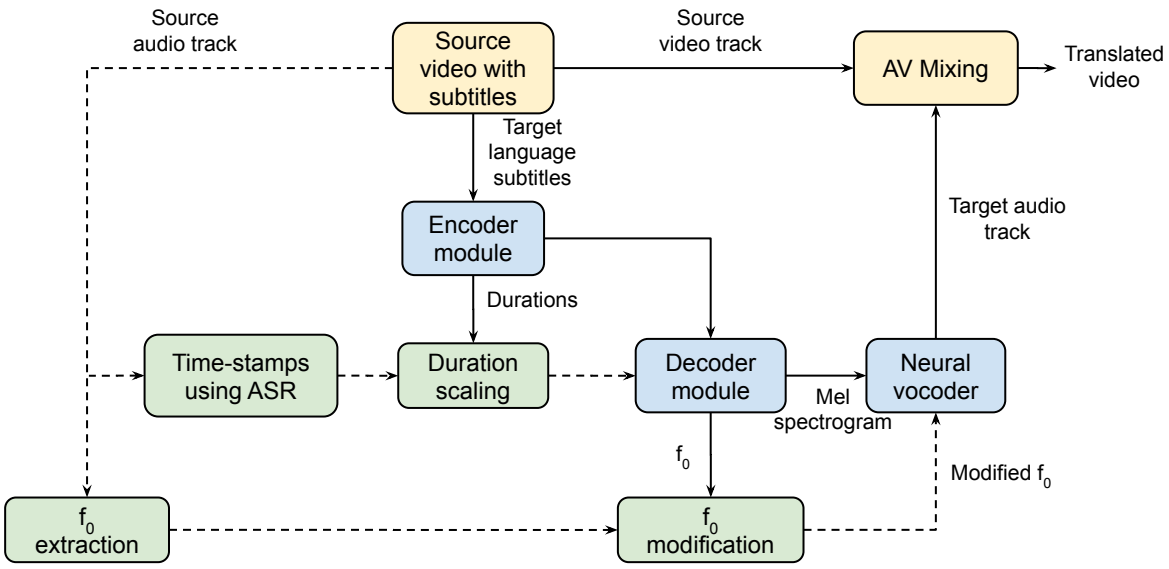

Fig. 1: Block diagram of the proposed subtitle synthesis method. Prosody modification blocks are coloured in Green and Prosody-TTS system modules are coloured in Blue.

the same duration. Energy modification helps in emphasizing certain words.

The salient features of the proposed method are:

- Subtitles can be synthesized in any foreign language, provided the Text-to-speech synthesis system of the corresponding language.
- The prosody of the synthesized speech can be modified to match the prosody of the source video.
- This is the first subtitle synthesis approach proposed for synthesizing an audio track in the desired language from the subtitles of a video so that the video can be viewed in the desired language.

The rest of the paper is organized as follows. The proposed subtitle synthesis method and the prosody modification are explained in section II. The information about the experimental setup and the issues to be addressed in the subtitle synthesis approaches are detailed in section III. Section IV describes the results of the proposed system, and section V concludes the paper with future work.

II. Proposed subtitle synthesis method

The block diagram of the proposed subtitle synthesis system is shown in the figure 1. The original video recorded in the native language is referred to as the source video. The audio track and the video track are separated from the source video, which are referred to as the source audio track and the source video track, respectively. Subtitles are provided for the foreign languages, and we consider the desired target language subtitles. This target language subtitle text is fed to the encoder module of the Prosody-TTS system, which predicts the duration of each phone in the given subtitles.

The bottleneck representations of the encoder module are fed to the decoder module of the Prosody-TTS system, which predicts the fundamental frequency (f_0) and uses it for conditioning the melspectrogram prediction. The predicted duration of the phones and the fundamental frequency are modified according to the prosodic parameters extracted from the source audio track to incorporate the source prosody in the target audio track. The predicted melspectrogram and the modified fundamental frequency are given as the input to the non-autoregressive neural vocoder, which generates the target audio track. Finally, the source video track and the target audio track are mixed using the audio-video mixing to get the translated video in the target language.

When the subtitles are not available in the target language, we extract them from the source audio track, as shown in figure 2. The source language Automatic speech recognition (ASR) system is used to transcribe the source audio track into the source subtitle text. A machine translation (MT) system is then used to translate the source subtitle text into the target subtitle text and fed to the encoder module of the Prosody-TTS system. This case is similar to the Speech-to-Speech translation task.

A. Prosody-TTS system

Text-to-Speech synthesis system has significant applications in human-machine interactions, speech-to-speech translation, etc. The speech samples should be synthesized with the desired prosody and emotion in these applications. The fundamental frequency, duration, and energy are the parameters responsible for the prosody and emotion in the synthesized speech. So, we need to have control of these prosodic parameters to control the prosody and the emotion. Most of the state-of-the-art speech synthesis systems like Tacotron [1] don't have

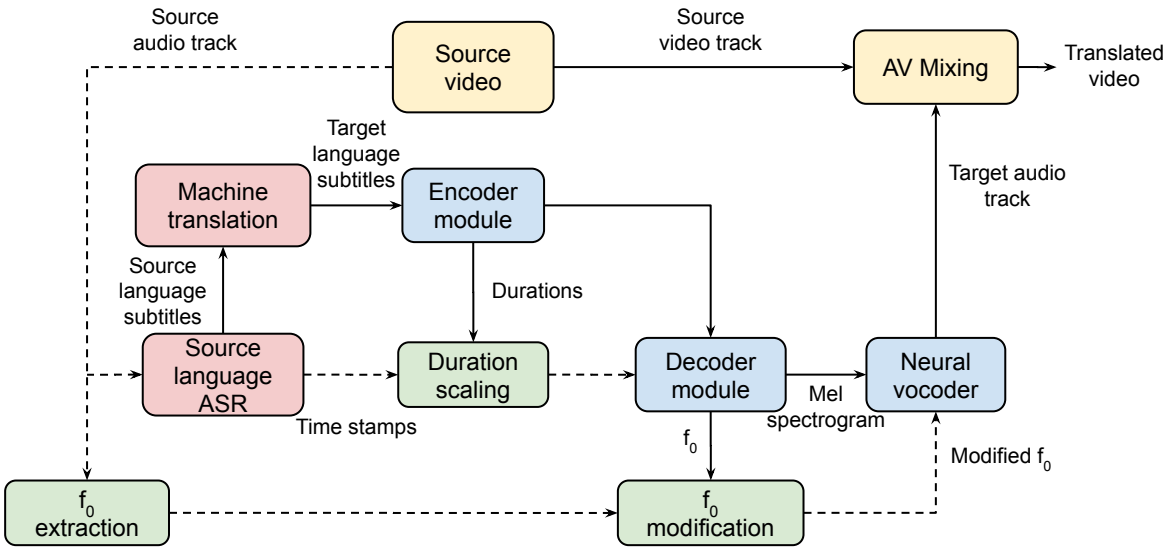

Fig. 2: Block diagram of the proposed subtitle synthesis method when the subtitles are not available in the target language. Prosody modification blocks are coloured in Green and Prosody-TTS system modules are coloured in Blue. Subtitle generation modules are coloured in Red

the provision to control these prosodic parameters, so the prosody of the synthesized speech cannot be modified in these systems. The Prosody-TTS system [8] can modify the prosodic parameters and synthesize the speech from the text in the desired prosody, which makes the speech synthesis system useful for expressive speech synthesis applications. So, we use the Prosody-TTS system in the proposed subtitle synthesis method, where we need to synthesize the speech in the desired prosody.

The Prosody-TTS system consists of an encoder module, a decoder module, and a neural vocoder. The subtitle text sequence of the target language $X = \{x_1, x_2, \ldots, x_N\}$ is converted into character embeddings $C = \{c_1, c_2, \ldots, c_N\}$ and fed to the encoder module. N is the total number of phonemes in the given target language subtitle text. Each character embedding c_n is the 128-dimensional fixed non-learnable embedding sampled from the standard distribution corresponding to the phoneme x_n. The encoder module predicts the total duration sequence $T = \{t_1, t_2, \ldots, t_N\}$ by maximizing the probability function $p(t|c)$, where t_n is the duration of the phoneme x_n in milliseconds (ms).

The bottleneck representations of the encoder module are upsampled using the predicted durations to the acoustic frame rate as they are at linguistic frame rate. These upsampled representations are given to the decoder module to predict the fundamental frequency and the melspectrogram. An excitation signal is created with the predicted (f_0) and provided as an input to the neural vocoder along with the predicted melspectrogram. The vocoder is a source-filter model-based non-autoregressive neural vocoder, which predicts the target audio track from the given excitation signal and the melspectrogram. The durations and the fundamental frequency are modified according to the prosody of the source audio track to incorporate the source prosody.

1) Duration modification: The duration of the source audio track and the target language audio track should be the same to incorporate the target audio track in the source video. So, the predicted durations from the encoder module of the Prosody-TTS system are modified according to the timestamps obtained from the source audio track using the ASR system of the source language. The phrases of the source and target audio tracks should be roughly of the same durations. The predicted duration of the silence regions is modified according to the silence regions of the source audio so that they sync with the source video. We observed that the longer inter-sentence silences are aligned well, but it is difficult to align the shorter intra-sentence pauses. So, as future work, we will find a way to automatically align the pauses using the techniques like dynamic time warping (DTW), etc.

2) f_0 modification: The fundamental frequency is modified to ensure that it matches the prosody and the emotion in the source audio. The source f_0 track is extracted from the source audio track using the speech analysis tools like WORLD [9], REAPER [10], etc. Deviations from the mean fundamental frequency are calculated from the source f_0 track. These deviations are incorporated in the predicted fundamental frequency from the decoder module of the Prosody-TTS system. The source, predicted, and the modified f_0 tracks are shown in the figure 3. The modified fundamental frequency is given as input to the neural vocoder to create the excitation signal.

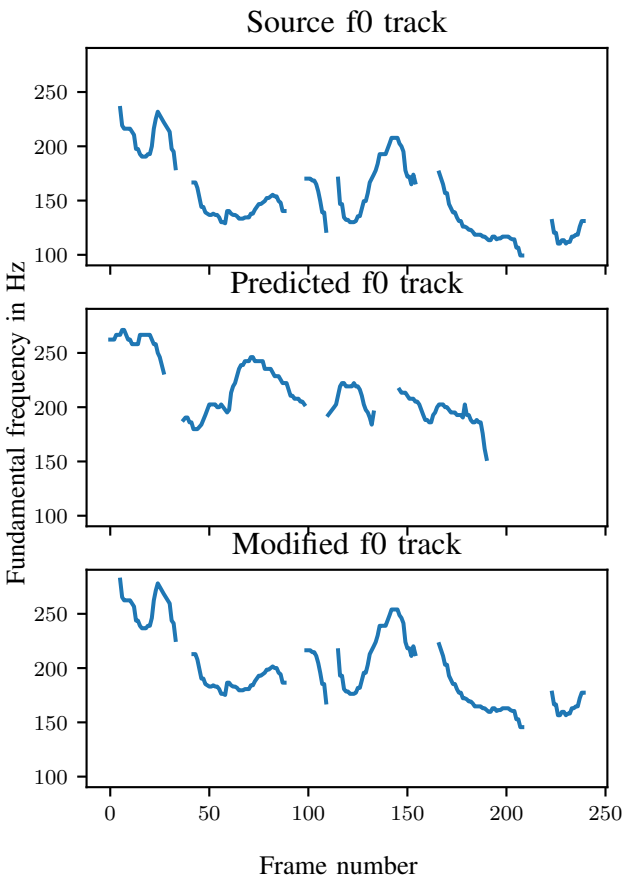

Fig. 3: f_0 track of the source audio track (Top). Predicted f_0 track from the decoder module (Middle). Modified f_0 track (Bottom).

III. Experimental setup

The major blocks of the subtitle synthesis system are the modules of the Prosody-TTS system. We need to train a Prosody-TTS system in the target language. In this paper, we used single speaker video data, like lecture videos, to translate them into the target language. So, we trained the Prosody-TTS system with the single speaker data in the target language. This can be extended to multi-speaker video data, where we need to train a multi-speaker Prosody-TTS system by conditioning on the speaker information. But, we need to get the speech boundaries of the speakers from the source audio track using Speaker diarisation, which gives the information about "Who spoke when?".

In this paper, we have considered English-Telugu and Hindi-Telugu video translation, where we synthesize the Telugu language subtitles and incorporate the synthesized audio tracks in the source videos. We have considered both the cases of having the subtitles and without having the subtitles in the source videos. For English-Telugu translation, subtitles are available in the English language for the English lecture video. So, we have used Google Translate to translate the English subtitles into Telugu subtitles. In the Hindi-Telugu translation task, subtitles are not available for the Hindi lecture video. So, we used the Automatic Speech Recognition (ASR) [] system to transcribe the Hindi audio track into the Hindi text and then used Google Translate to translate the Hindi text to Telugu text. So, for both tasks, we need to train a Prosody-TTS system in the Telugu language.

A. Prosody-TTS training

The Prosody-TTS system is trained on the Telugu language data from the IndicTTS database [11]. Since the English and Hindi source videos are from a male speaker, we have used Telugu male speaker data to train the TTS system. The database has around 2400 utterances, which correspond to 4 hours. It has text and the corresponding speech file, which is sampled at 16 kHz. First, the duration of the phones is obtained using the HTK forced alignment [12]. WORLD vocoder [9] is used to extract the fundamental frequency from the speech files with a frame length of 25 ms and a frame-shift of 5 ms. 80-dimensional Melspectrograms are extracted from the speech files with a frame-length of 25 ms, frame-shift of 5 ms, and 2048 point FFT. The Prosody-TTS system is then trained to synthesize the target Telugu audio track from the given input Telugu subtitles.

B. Subtitle Synthesis

In this paper, we have considered the NPTEL lecture videos[1] to translate them into the other desired languages using the subtitle synthesis approach. These are single speaker videos recorded in native languages. First, the audio track and the video track are separated from the source videos using the FFmpeg tool [13]. Then, we extract the Telugu language subtitles for the English-Telugu and Hindi-Telugu tasks. The extracted Telugu subtitles are fed to the encoder module of the Prosody-TTS system, which predicts the duration of the phones. ASR system is trained in the source language to give the timestamps of the source audio track. The ASR acoustic model was based on purely sequential trained TDNN with the lattice-free maximum mutual information (LF-MMI) objective [14], and it was trained using Kaldi [15] with 40 hours of labeled data. The input features to the TDNN were 40-dimensional Mel-frequency cepstral coefficients (MFCCs). The force-aligned timestamps from this ASR system are used to align the inter-phase silences by modifying the predicted durations. The decoder module predicts the fundamental frequency and the melspectrogram. The source f_0 track is extracted from the source audio track using the WORLD vocoder or the REAPER algorithm. The extracted source f_0 track is used for modifying the predicted f_0 track by using the deviations from the mean f_0. The neural vocoder takes the modified f_0, the melspectrogram, and generates the target audio track. FFmpeg tool is used for mixing the target audio track with the source video track.

[1] https://nptel.ac.in/

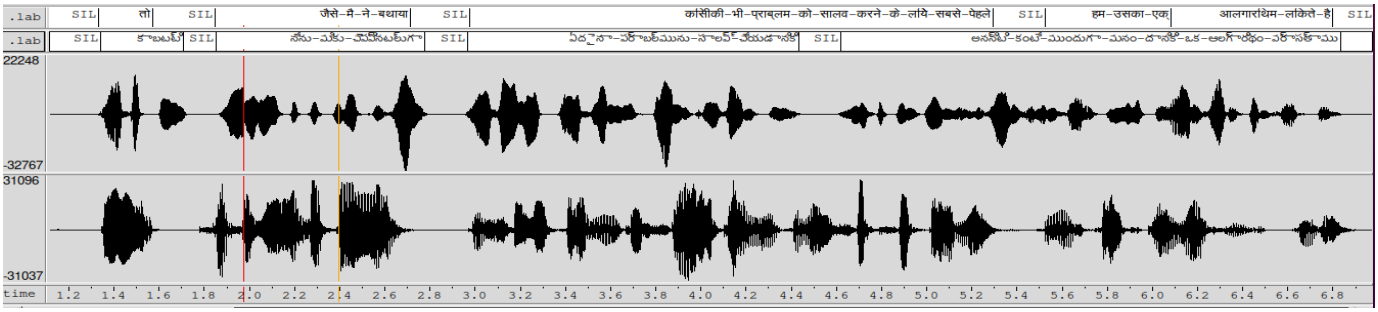

Fig. 4: Source (Top) and Target (Bottom) audio tracks of Hindi-Telugu translation with the word level labels

C. Issues to be addressed

The subtitle synthesis approach has several issues to be addressed to improve the quality of experience. Although we can synthesize the target language audio track with the desired prosody, it is very difficult to align the synthesized audio track with the source video. The silence regions in the target audio track should sync with the source video. The duration of the audio track should match with the frames of the source video to mix them. The synthesized audio track should match the events in the video. The inter-sentence silences are aligned using the timestamps from the ASR system and modifying the predicted durations. But, it is difficult to align the intra-sentence pauses.

We modify the fundamental frequency to incorporate the desired prosody and emotion in the target audio by calculating the deviations from the mean f_0. Although incorporating the f_0 deviation is good for matching the prosody and the emotion in the source video, the unintended words are getting emphasized. Energy modification helps in emphasizing certain words. But, it is difficult to incorporate as it needs inputs from the ASR and the Machine translation systems to identify which words are to be emphasized.

The quality of the experience can be improved by using image processing algorithms [16] to lip-sync the video with the translated audio. In some of the cases, the source speaker characteristics are needed to be preserved, which can be done by using voice conversion algorithms [17]. In the multi-speaker video scenario, we need to get inputs from the Speaker diarisation algorithm about the timestamps of who spoke when. In this case, we need to have a multi-speaker TTS system that can generate different characteristics when conditioned on the speaker information. We have to use voice activity detection algorithms to separate the voice and music if they are not recorded in the separate channels. With these issues, the subtitle synthesis approach gives vast areas for the research groups to improve the quality of experience.

IV. Results

The quality and the naturalness of the Hindi-Telugu and English-Telugu translated videos are evaluated using the Mean opinion scores (MOS) from the native Telugu language speakers. We have considered 10 videos for the Hindi-Telugu translation task and 10 videos for the English-Telugu task to evaluate the performance of the proposed subtitle synthesis approach. The subtitles from these 20 videos are fed to the Prosody-TTS system to generate the Telugu audio tracks and mix them with the source videos. Around 30 native Telugu language speakers are considered for evaluating the translated videos. Reviewers were asked not to consider the lip-sync differences between the audio and the video.

Subtitles are not available for the Hindi videos, where we use ASR and MT to extract the subtitles from the source audio. English subtitles are available for the English videos, which are manually translated to the Telugu subtitles to suppress the error from the MT system in the evaluation process. MOS are evaluated separately for the English-Telugu and Hindi-Telugu translations since the errors from the ASR and MT systems contribute to the quality of the translated video in the Hindi-Telugu translation task. Table I shows the obtained MOS with a 95% confidence interval for both the translation tasks. The quality and the naturalness can be improved by addressing the issues discussed in section III-C. The waveforms of the source and target audio tracks of the Hindi-Telugu translation, along with the word level labels, are shown in the figure 4. We can observe that the inter-sentence silences are aligned well, but the intra-sentence pauses are not aligned correctly.

TABLE I: The MOS of the source and the translated videos with a 95% confidence interval

Task	MOS
Source videos	4.128 ± 0.14
Hindi-Telugu	3.723 ± 0.11
English-Telugu	3.846 ± 0.08

V. Conclusion and Future work

In this paper, we proposed a preliminary approach for synthesizing subtitles of a video in the target language and incorporating the synthesized audio track in the source video. The prosody and the emotion of the synthesized audio track should match with the source video. So, we used the Prosody-TTS system to generate audio tracks from the subtitles, which has provision to modify the prosodic parameters like fundamental frequency, duration, etc. Although the translated videos are quite natural, there are several issues that are to

be addressed in the subtitle synthesis approach. As a future work, we will address these issues to improve the quality of the experience. Image processing algorithms can be used to lip-sync the video to enhance the naturalness further. In the future, subtitles can be replaced with auto-created audio tracks, which have numerous applications in the education and the entertainment sector.

REFERENCES

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[2] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895

[3] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," *CoRR*, vol. abs/1702.07825, 2017. [Online]. Available: http://arxiv.org/abs/1702.07825

[4] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *ICLR*, 2017.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[6] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *CoRR*, vol. abs/1802.08435, 2018. [Online]. Available: http://arxiv.org/abs/1802.08435

[7] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," 2019.

[8] P. Giridhar and K. S. R. Murty, "Prosody-tts: A non-autoregressive end-to-end speech synthesis system with prosody modification," 2021, Manuscript submitted for publication.

[9] M. MORISE, F. YOKOMORI, and K. OZAWA, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[10] D. Talkin, "Reaper: Robust epoch and pitch estimator," https://github.com/google/REAPER, 2015.

[11] A. Baby, A. Thomas, N. L, and T. Consortium, *Resources for Indian languages*. Community-Based Building of Language Resources, 09 2016.

[12] S. J. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. L. Moore, J. J. Odell, D. Ollason, D. Povey, Valtchev, and P. C. Woodland, "The htk book version 3.4," 2006.

[13] FFmpeg Developers., "ffmpeg tool." [Online]. Available: http://ffmpeg.org/

[14] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[16] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[17] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.