ANALYSIS OF CONVERSATIONAL SPEECH WITH APPLICATION TO VOICE ADAPTATION

Bhagyashree Mukherjee[1] *, Anusha Prakash*[2]*, Hema A. Murthy*[1]

[1]Department of Computer Science & Engineering, Indian Institute of Technology Madras, India
[2]Department of Electrical Engineering, Indian Institute of Technology Madras, India

bhagyashree@cse.iitm.ac.in, anushaprakash@smail.iitm.ac.in, hema@cse.iitm.ac.in

ABSTRACT

Conversational speech has always been challenging in the context of text-to-speech synthesis (TTS). Most speech synthesis systems are trained on read speech data recorded in a studio environment. But, the intelligibility of TTS systems degrades drastically when using conversational speech. The proposed work attempts to perform extensive analysis on the issues in dealing with conversational speech compared to read speech. As an application, we try to dub the lectures available in English into an Indian language (Hindi) in the original speaker's voice. The task is difficult as classroom lectures are extempore, with variations in speaking rate, and contain speaker mannerisms that lead to disfluencies. We analyze the capability of end-to-end TTS systems in modeling lecture-based data. Based on the analysis, an attempt is made to adapt "read speech TTS system" using conversational speech data to produce lectures in the original speaker's voice.

Index Terms— conversational speech, cross-lingual, voice adaptation

1. INTRODUCTION

The rapid drift towards online learning platforms has led to the demand for lecture videos in regional languages. A typical dubbing system involves recognizing the original speech, translating the recognized text into the required native language, synthesizing the translated text and syncing the synthesized audio with the original video. To enhance understanding and present a real-time scenario even in the dubbed lectures, synthesizing the translated text in the original speaker's voice provides an extra edge. The ultimate goal of this work is to perform cross-lingual voice adaptation for classroom lectures, which is essentially conversational in nature. Training text-to-speech (TTS) systems for conversational speech is challenging as classroom lectures are extempore, with variations in speaker mannerisms and speaking rate. We analyze the differences across read speech (studio-recorded) and conversational speech (classroom lectures), both at the training level and at the synthesized output. We then perform voice adaptation for conversational speech in an end-to-end framework, such that they are capable of synthesizing bilingual text.

In the context of conversational speech, several techniques are proposed for automatic speech recognition (ASR), and segmentation [1, 2, 3, 4]. [5] and [6] present the idea of speaker adaptation to reduce the word error rate for ASR. Disfluencies are inevitable in conversations. Several papers, such as [7, 8] perform disfluency analysis and detection in speech recognition. Analysis based on pauses, utterance length, and vowel duration is performed in [9]. In the TTS domain, [10, 11, 12] perform experiments on spontaneous speech synthesis and evaluations. For voice adaptation and conversion, techniques such as speaker disentanglement [13] to separate the speaker and content representations, speaker adaptation [14, 15, 16, 17, 18] one-shot voice conversion [19] for unseen speakers, and various other encoder-decoder architectures are developed to achieve human-like speech. Cross-lingual voice conversion techniques include source-content separation and phonetic-posteriorgrams as mentioned in [20, 21, 22, 23, 24]. To the best of our knowledge, the current work is the very first attempt to analyze the underlying attributes of conversational speech and perform cross-lingual voice adaptation for lectures in an E2E framework. The current work is set in the context of low resource data.

The authors in [16] present the idea of generic TTS systems for Indian languages and state that the target voice is preserved even with as little as 7 minutes of adaptation data. Inspired by this idea in the context of voice adaptation, we adopt a similar approach for conversational speech adaptation. It is observed that there is a significant degradation in the synthesis quality and intelligibility when the same amount of lecture-based audio is used for adaptation. This is primarily because the data used in [16] is "read speech" recorded in a professional environment, whereas the current work uses conversational speech data. We attempt to understand the reason for this degradation by analyzing the variations in lecture-based speech with respect to read speech, thereby trying to adapt conversational speech from read speech system to obtain the original speaker's voice.

As part of this work, we train four E2E TTS systems–

System 1: trained with only conversational speech data (English), *System 2:* trained with only read speech data (Hindi+English), *System 3:* bilingual read speech TTS system adapted to manually cleaned conversation speech data (English), *System 4:* bilingual read speech TTS system adapted to automatically cleaned conversation speech data (English). Of these systems, the adapted systems (System 3 and 4) are the proposed systems for voice adaptation. For System 4, the conversational speech data is automatically curated by considering only the text-audio pairs that have a high confidence score with an ASR system [25]. Analysis of synthesized audio of these systems indicates that improper sentence endings and disfluencies affect the prediction of attention and sentence boundaries, in turn affecting the TTS quality.

The rest of the paper is organized as follows: Section 2 analyzes various attributes across read speech and impromptu classroom lectures and highlights the complications in using the latter. Extensive analysis is carried out on pitch variations, syllable rates, and disfluencies within the lectures. Section 3 details the basic system architecture used for training and adaptation. Section 4 analyzes the synthesized outputs of the trained systems and discusses the results of subjective and objective evaluations. The work is concluded in Section 5.

2. CHALLENGES IN CONVERSATIONAL SPEECH

E2E speech synthesis systems have become state-of-the-art in generating intelligible and high-quality audios. But the exceptional quality of voice is limited to models trained on speech recorded by professionals in a studio environment. System building becomes more challenging when we deal with conversational speech. Unlike recordings of read speech in a studio environment, conversational speech may be recorded in noisy environments, such as classrooms, airports, hospitals, etc. In this paper, we focus on lectures from National Programme on Technology Enhanced Learning (NPTEL) [26] recorded in classroom environments and analyze the issues in dealing with these audios.

2.1. Dataset Description

NPTEL is an online educational platform [26] which has 52000+ hours of transcribed content and more than 51000 hours of subtitled videos. Some of the transcriptions are generated using an ASR system and converted into a SubRip Subtitle (SRT) file based on video timestamps. For the analysis task, we have considered English speech data of two professional native Hindi speakers from IndicTTS database [27], which is read speech, and data of two NPTEL faculty [26], which is conversational speech. One male and one female speaker is considered from each set. Each dataset is of approximately 1 hour in duration. The datasets used as a part of this work will be made available on request.

2.2. Pitch:

Pitch is one of the most critical parameters for determining speaker identity [28]. For any voice manipulation task, modeling the pitch is essential to producing the target voice. Pitch contours help modify prosodic features, and several analyses have been carried out in the literature for expressive speech [29]. We compare and analyze the pitch histograms across read speech recordings and conversational speech of different speakers. From Figure 1, it is observed that lectures have a wide variation in pitch compared to read speech. Since the pitch is an essential parameter in speech, the broad pitch range takes a toll on intelligibility and speaker characteristics in the synthesized audio.

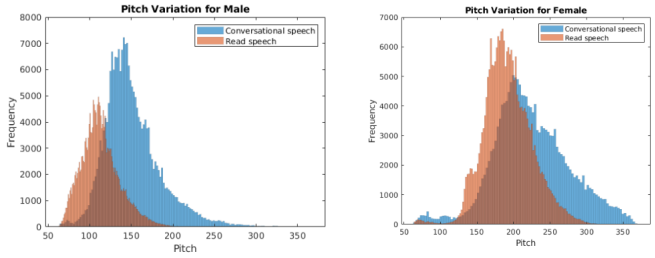


Fig. 1. Pitch variations across read and conversational speech

2.3. Syllable rate:

Syllable rate is an estimate of the number of syllables uttered per second. To build a TTS model, the training data is expected to have a uniform syllable rate across all the utterances. This ensures that the basic sound units are modeled robustly during training. In the absence of a constant syllable rate, the quality of TTS output becomes inconsistent. Varying syllable rate is an important characteristic of conversational speech. As seen from Figure 2, there is a significant fluctuation in the syllable rate across the conversation speech lectures. In contrast, the syllable rate remains comparatively constant for read speech.

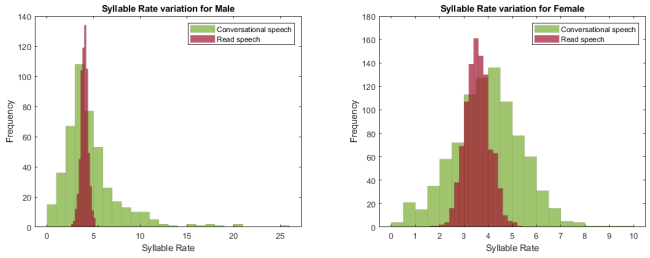


Fig. 2. Syllable rate variations across read and conversational speech

2.4. Disfluency rate:

Disfluencies are unavoidable attributes in spontaneous speech. Classroom lectures are unscripted; hence lots of disfluencies like *umm*, *ah*, *okay*, *so*, *right*, *is it*, etc., are common during a lecture. It may not always be reflected in the text, leading to a mismatch between the audio and corresponding transcriptions. Disfluency analysis is performed on one male and female lecture for conversational speech. About 30 minutes of the course is considered for analysis. Disfluency rate is defined as the number of disfluencies as a percentage of the total number of words. As seen from Table 1, the average disfluency rate of conversational speakers is approximately 4%. In read speech, there are no disfluencies as recordings are performed carefully.

Table 1. No. of disfluencies in a 30 minute lecture audio (conversational speech)

Speaker	**No. of disfluencies**	**Total no. of words**	**Disfluency rate**
Female	194	4750	4.08%
Male	162	4920	3.29%

3. SYSTEM DESCRIPTION

Four E2E TTS systems are trained as part of this work. As seen from Figure 3, Systems 1 and 2 are directly trained, while Systems 3 and 4 are adapted TTS systems. The basic modules involved in training TTS systems are described first, followed by a detailed description of each of the four systems. For conversational speech data, speech waveforms are extracted from the NPTEL classroom lectures of only one female speaker [26], which is the target speaker. It is to be noted that the NPTEL lectures are in English. For read speech data, Hindi and English recordings of a professional Hindi female speaker are considered (IndicTTS database) [27].

3.1. Modules involved in training

E2E systems are directly trained using <text, audio> pairs. First, the transcriptions are converted into corresponding phone-based representations, and a Tacotron2 based network is trained [30]. The Tacotron2 network takes text as input and generates the mel-spectrogram. The Waveglow vocoder then reconstructs speech from the mel-spectrogram [31].

3.1.1. Phone-level representation

We follow the procedure in [32] to convert text to its phone-based representation. Hindi and English words are handled separately. Hindi words are parsed through a unified parser for Indian languages [33]. The parsed output is in terms of the phone-based common label set (CLS) representation [34]. English words in the training data are parsed using a manually prepared dictionary due to the lack of a good Indian English phone-based parser. The dictionary provides word-to-CLS conversion. To handle out of vocabulary (OOV) English words, a classification and regression tree (CART) is built using the dictionary.

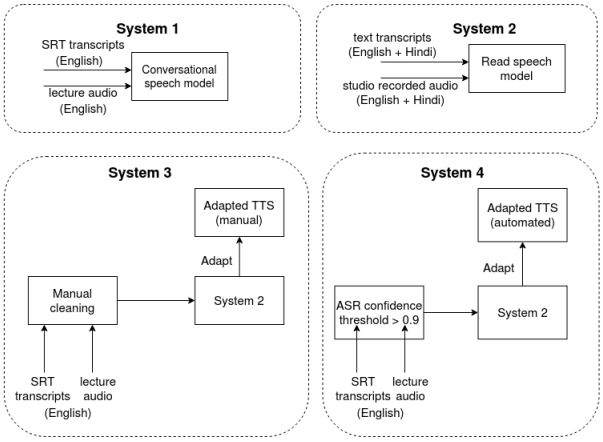


Fig. 3. Block diagram of the four TTS systems

3.1.2. Training

ESPnet's implementation [35] of Tacotron2 [30] is used for training the TTS systems. Tacotron2 has an encoder-decoder architecture that uses an attention mechanism. Only text-audio pairs are needed for training. The phonetic dictionary maps the text in the form of tokens. These tokens are fed into the encoder model and converted into fixed-length vectors. The fixed-length vectors are passed to the decoder for the prediction of mel-spectrogram for each frame. x-vectors are used as speaker embeddings in the training stage. x-vectors are deep neural network (DNN) based embeddings that capture speaker-specific information [36]. They are extracted from a pre-trained model using Voxceleb data. The key idea in using x-vectors is to exploit this latent speaker information for the voice adaptation task.

3.1.3. Vocoder

Waveglow [31] is used as the vocoder to reconstruct the speech waveform. It is a generative model that samples from a simple distribution to model the audio distribution conditioned on the mel-spectrogram. It uses maximum likelihood estimation as a training cost function. The pre-trained LJSpeech waveglow model is used as an initial model and retrained using the target speaker's data. The lecture recordings are subdivided into smaller chunks, each having a duration of 15 seconds, and used for training the waveglow network.

3.1.4. Adaptation

The bilingual Tacotron2 model parameters are fine-tuned on the adaptation data of the target speaker. x-vectors are extracted for the target speaker and used for training the adapted TTS. This helps in fine-tuning the model parameters specific to the target speaker. It is to be noted that the original network is trained on read speech data, while the adaptation data is conversational in nature.

3.2. System 1: TTS trained on conversational speech

In order to achieve our objective of generating lectures in the original speaker's voice, the most basic approach is to build a TTS system by pooling the available speaker's data. The english text is converted into phone-level representation as described in Section 3.1.1. This is our conversational speech model. But there are a few challenges associated with it. The availability of correct transcripts is highly dependent on the performance of ASR. Secondly, the audio segmentation is based on voice activity detection (VAD), which may or may not end in a meaningful sentence. Hence, the E2E system struggles to model the context and end of a sentence. Although this system produces speech in the target speaker's voice, it may not be viable for cross-lingual synthesis. The phonotactic variations between English and Indian languages make it difficult to synthesize Indian language text using an English TTS. The speaker's data is also not available in any Indian language. System 1 is trained with 12 hours of lecture data (English) of the target speaker.

3.3. System 2: Bilingual TTS trained on read speech

We attempt to build a phone-based Hindi+English model capable of handling bilingual text. It is to be noted that monolingual Hindi and English recordings of the same speaker are combined during training. This model is built on read speech data recorded in a studio environment, and each audio is a complete sentence enabling the E2E model to learn robustly. 8.5 hours of Hindi and an equal amount of English data of one native female speaker from Indic TTS [27] is used to build the model. Only for System 2, the Waveglow vocoder is trained using the read speech data, as opposed to conversational speech data used in the other systems. System 2 cannot synthesize speech in the target speaker's voice because System 2 is a single speaker TTS.

3.4. System 3: Bilingual read speech TTS adapted to manually cleaned lecture data

We adapt the bilingual model (System 2) using manually cleaned 25 minutes of the speaker's data to generate translated audio in the target speaker's voice. The manually curated data is carefully chosen where hesitations, pauses, and other disfluencies are removed, sentences are complete, and syllable and pitch rate variance are maintained similar to that of read speech systems. This allows learning the model parameters to be learnt more robustly.

3.5. System 4: Bilingual read speech TTS adapted to automatically curated lecture data

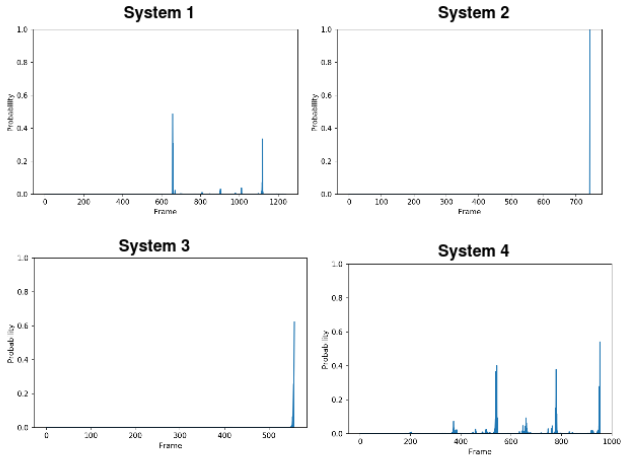
The manual cleaning of the transcriptions in System 3 requires a lot of effort and is time-consuming. In an attempt to automate this step, we used an ASR model trained on 180 hours consisting of read speech and conversational speech using the Kaldi chain model framework [37]. The ASR system achieves a word error rate (WER) of 6.20%. Using this ASR model, we compute the confidence score of each audio. The assumption is that we may eliminate the erroneous audio-transcription pairs as well as the undetected disfluencies by using a higher confidence threshold. Using a threshold of 0.9 confidence on the training data used for System 1, we obtain 40 minutes of data for adapting bilingual read speech TTS (System 2).

4. ANALYSIS OF SYNTHESIZED SPEECH

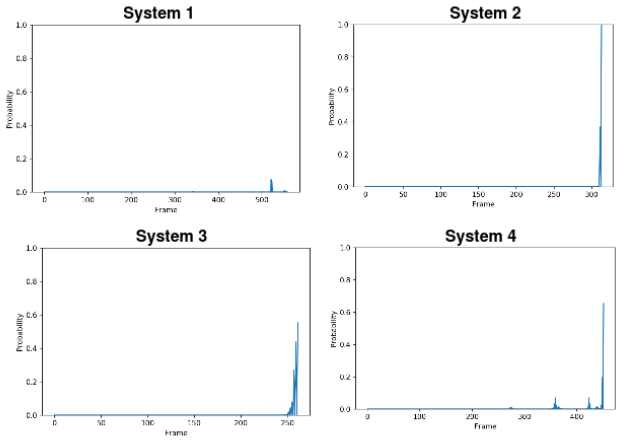
The synthesized output generated by the four TTS systems are evaluated in terms of qualitative analysis– prediction of the end of sentence and attention. Further, subjective and objective evaluations– degraded mean opinion score (DMOS), speaker similarity, and mean cepstral distortion (MCD) are discussed below.

4.1. Qualitative Analysis

Stop token or end of sentence token (<eos>) is predicted during inference time at the frame level. Systems 1 and 4, which are trained/adapted using conversational speech data, fail to model the end of a sentence. An improper sentence ending in Systems 1 and 4 may result in an incomplete utterance or random words or a disfluency towards the end of the synthesized audio. This is mainly because the transcriptions may not always be complete sentences, which affects the system training. The manual curation of data in System 3 helps to predict sentence boundaries accurately. We show two cases in Figure 4. In Case 1, the sentence can have multiple endings; hence, Systems 1 and 4 get confused and predict multiple frames to be <eos> with a small probability. In contrast, the read speech system (System 2) predicts the sentence end with a probability of almost 1. In System 3, we see that the manual cleaning of the lectures helps predict the sentence end correctly. This is because, in manual cleaning, we ensure sentences are complete and meaningful. In Case 2, System 1 fails to predict <eos>, whereas Systems 2, 3, and 4 correctly predict sentence boundary. This peculiarity in System 4 is because it is adapted from System 2 (which is trained on read



Case 1 : *" But that's where we are going so by the end of the course we would understand the technicality that drive the four hundred gigabits per second "*



Case 2 : *" And now we are going to talk about four hundred gigabits per second "*

Fig. 4. Prediction of end of sentence for different systems

speech data) and is hence able to predict boundaries correctly in some cases.

A similar trend is seen even in the attention plots for a sample utterance, as shown in Figure 6. Systems 1 and 4 struggle to predict the sentence's beginning and ending, whereas Systems 2 and 3 perform better. Although System 4 is adapted from read speech, the fine-tuning of the model using conversational speech data degrades the attention due to improper sentence ending as well as the presence of disfluencies. Further, we show a comparative analysis on focus rate of attention (FRA) [38] across 10 utterances synthesized using the four systems. We see that the FRA for Systems 1 and 4 is lower, whereas the FRA for manually cleaned System 3 is comparable to that of read speech System 2. This clearly demonstrates the fact that conversational TTS is challenging and the importance of complete sentences for training and adaptation. On carefully examining the adaptation data, we see that the pitch and syllable rate distribution show a higher variance in case of System 4 and a smaller variance in System

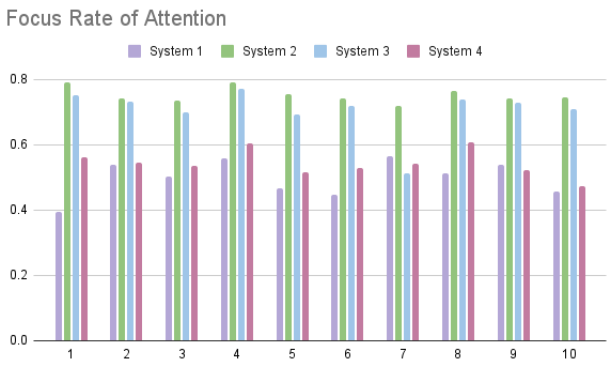


Fig. 5. Focus rate of attention across various systems for 10 sample utterances

3. This variation in the distributions can be another reason why System 4 is not as good as System 3 and highlights the difficulty in automating the process.

4.2. Subjective and objective evaluations

We perform subjective evaluations for monolingual and bilingual (Hindi+English) sentences using Systems 1, 3, and 4. System 2 is excluded as it doesn't generate speech in the target speaker's voice. In monolingual synthesis, test sentences are in English. The bilingual Hindi+English test set is a "cross-lingual" scenario, as the original speaker's data is in English, and we are attempting to synthesize bilingual sentences. Bilingual test sentences are obtained by translating the English transcriptions of the lectures into Hindi while retaining the technical terms in English. Two types of subjective evaluations are performed– degraded mean opinion score (DMOS) [39] and speaker similarity test.

In the DMOS test, listeners rate the quality of the synthesized speech, and the score is normalized with respect to that of the original speech. The DMOS rating is from 1-5, 5 being human-like quality. In speaker similarity, the listeners have to rate the speaker's similarity with respect to the target speaker (reference). DMOS and speaker similarity scores are presented in Table 2 and Table 3, respectively. Seventeen listeners participated in each evaluation and rated 25 sentences (7 from each system + 4 original) each for monolingual and cross-lingual tasks. For speaker similarity, 27 sentences were evaluated (8 from each system + 3 original).

Table 2. DMOS scores

System	monolingual	cross-lingual
System 1	3.41	2.49
System 3	3.89	3.97
System 4	2.82	2.87

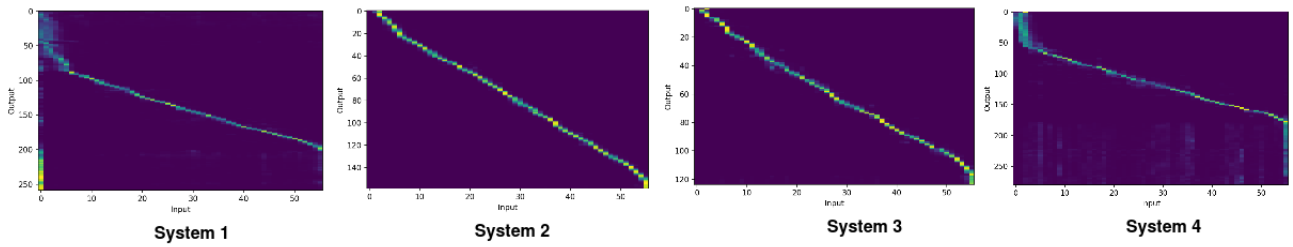


Fig. 6. Prediction of attention across various systems for a sample utterance

Table 3. Speaker similarity scores

System	monolingual	cross-lingual
System 1	3.81	1.92
System 3	3.64	3.63
System 4	2.58	2.32

Table 4. MCD scores on English utterances generated using Systems 1, 3 and 4

System	System 1	System3	System 4
MCD score	16.24	8.79	16.43

Objective evaluations across Systems 1, 3, and 4 are conducted. Synthesized utterances generated from the systems are compared with respect to the original audio using dynamically time-warped (DTW) mel-cepstral distortion (MCD) scores [40]. 20 unseen English sentences are considered for the MCD calculation (Table 4).

We clearly see that subjective and objective evaluations show System 3 (using manually curated data) to be the best. In System 1 (purely conversational TTS), although the monolingual DMOS and speaker similarity scores indicate good quality synthesis, the MCD score is also quite high (Table 4). The high MCD score reflects the distortions, repetitions, and added disfluencies present in the synthesized audio. This is also supported by qualitative analysis (Section 4.1). Cross-lingual DMOS and speaker similarity scores are low for System 1 because the entire phoneset of Hindi is not covered in English. System 4 gets a comparatively better score than System 1 in the cross-lingual context because of adaptation from the bilingual model. Still, repetitions and word skips are present in the synthesized audio, thus degrading the intelligibility and quality of the system.

On the informal evaluation of System 4, we see that there is a trade-off between the amount of conversational data used for adaptation and the system intelligibility for cross-lingual adaptation. As we increase the amount of conversational speech data, the pronunciation of Hindi words in the synthesis output degrades. This is mainly because the model gets biased towards English. We also observe some disfluencies inserted in the beginning and end of the synthesized audio. Although the results with manual cleaning seem promising, more sophisticated techniques like pruning based on syllable rate, pitch and <eos> have to be explored to automate cross-lingual voice adaptation using conversational speech data. The synthesized samples using the four systems can be found in the link:
www.iitm.ac.in/donlab/preview/ASRU2021/index.html

5. CONCLUSION

In this work, we have tried to understand the underlying peculiarities of conversational speech with respect to read speech. We have analyzed the performance of each of the trained E2E systems in terms of its synthesis quality and intelligibility. We have seen how the synthesized audio gets degraded when the utterances are not spliced properly. But with a more sophisticated technique for curating the data, it would be possible to build a good conversational TTS system in the target speaker's voice. This is clearly seen from the results of the proposed adapted system using the manually cleaned data. This is encouraging as it indicates that cross-lingual voice adaptation for conversational speech is possible, even in the low resource context. Further, we would like to explore the integration of pitch and syllable rate variation to achieve the target speaker's mannerisms.

6. ACKNOWLEDGEMENTS

This work was carried out as a part of the following projects, "Natural Language Translation Mission" (CS2021012MEIT003119) and "Speech to Speech Machine Translation" (CS2021152OPSA003119), funded by the Ministry of Electronics and Information Technology (MeitY), Office of the Principal Scientific Advisor (PSA) to the Government of India, respectively.

7. REFERENCES

[1] Marie Meteer and Rukmini Iyer, "Modeling conversational speech for speech recognition," in *Conference*

on Empirical Methods in Natural Language Processing, 1996.

[2] Ruchard Dufour, Vincent Jousse, Yannick Estève, Fréderic Béchet, and Georges Linarès, "Spontaneous speech characterization and detection in large audio database," *SPECOM, St. Petersburg*, 2009.

[3] Boros M et al, "Towards understanding spontaneous speech: Word accuracy vs. concept accuracy," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 2, pp. 1009–1012.

[4] Sadaoki Furui, "Recent advances in spontaneous speech recognition and understanding," in *ISCA & IEEE workshop on spontaneous speech processing and recognition*, 2003.

[5] Hiroaki Nanjo and Tatsuya Kawahara, "Unsupervised language model adaptation for lecture speech recognition," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[6] Takahiro Shinozaki and Sadaoki Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, vol. 1, pp. I–729.

[7] Vivek Rangarajan and Shrikanth Narayanan, "Analysis of disfluent repetitions in spontaneous speech recognition," in *2006 14th European Signal Processing Conference*. IEEE, 2006, pp. 1–5.

[8] Kourkounakis et al., "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6089–6093.

[9] Zue Victor et al, "The collection and preliminary analysis of a spontaneous speech database," in *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989.

[10] Shiva Sundaram and Shrikanth Narayanan, "An empirical text transformation method for spontaneous speech synthesizers," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson, "Spontaneous conversational speech synthesis from found data.," in *Interspeech*, 2019, pp. 4435–4439.

[12] Eva Székely, Jens Edlund, and Joakim Gustafson, "Augmented prompt selection for evaluation of spontaneous speech synthesis," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6368–6374.

[13] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *Proc. Interspeech 2018*, pp. 501–505, 2018.

[14] Kiyohiro Shikano, Satoshi Nakamura, and Masanobu Abe, "Speaker adaptation and voice conversion by codebook mapping," in *1991 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1991, pp. 594–597.

[15] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 952–963, 2006.

[16] Anusha Prakash and Hema A Murthy, "Generic indic text-to-speech synthesisers with rapid adaptation in an end-to-end framework," *Proc. Interspeech 2020*, pp. 2962–2966, 2020.

[17] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Tie-Yan Liu, et al., "Adaspeech: Adaptive text to speech for custom voice," in *International Conference on Learning Representations*, 2020.

[18] Yuzi Yan, Xu Tan, Bohan Li, Tao Qin, Sheng Zhao, Yuan Shen, and Tie-Yan Liu, "Adaspeech 2: Adaptive text to speech with untranscribed data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6613–6617.

[19] Ju-chieh Chou and Hung-Yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *Proc. Interspeech 2019*, pp. 664–668, 2019.

[20] Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma, "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," *Proc. Interspeech 2020*, pp. 2927–2931, 2020.

[21] Yi Zhou, Xiaohai Tian, Haihua Xu, Rohan Kumar Das, and Haizhou Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6790–6794.

[22] Berrak Sisman, Mingyang Zhang, Minghui Dong, and Haizhou Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 144–151.

[23] Daniel Erro and Asunción Moreno, "Frame alignment method for cross-lingual voice conversion," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[24] Seyed Hamidreza Mohammadi and Taehwan Kim, "Investigation of using disentangled and interpretable representations for one-shot cross-lingual voice conversion," *Proc. Interspeech 2018*, pp. 2833–2837, 2018.

[25] Laurent Besacier, Benjamin Lecouteux, Ngoc-Quang Luong, Kaing Hour, and Marwa Hadj Salah, "Word confidence estimation for speech translation," in *International Workshop on Spoken Language Translation*, 2014.

[26] "National programme on technology enhanced learning," https://nptel.ac.in/.

[27] Arun Baby, Anju Leela Thomas, Nishanthi N L, and Hema A Murthy, "Resources for Indian languages," in *Community-based Building of Language Resources (International Conference on Text, Speech and Dialogue)*, 2016, pp. 37–43.

[28] Anisha Yathigiri, Meenalatha Bathula, Susmitha Kothapalli, Susmitha Vekkot, and Shikha Tripathi, "Voice transformation using pitch and spectral mapping," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 1540–1544.

[29] R. S. Deo and P. S. Deshpande, "Pitch contour modelling and modification for expressive marathi speech synthesis," in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014, pp. 2455–2458.

[30] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[31] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[32] Anusha Prakash, A Leela Thomas, S Umesh, and Hema A Murthy, "Building multilingual end-to-end speech synthesisers for indian languages," in *Proc. of 10th ISCA Speech Synthesis Workshop (SSW'10)*, 2019, pp. 194–199.

[33] Arun Baby, NL Nishanthi, Anju Leela Thomas, and Hema A Murthy, "A unified parser for developing indian language text to speech synthesizers," in *International Conference on Text, Speech, and Dialogue*. Springer, 2016, pp. 514–521.

[34] Ramani et al., "A common attribute based unified HTS framework for speech synthesis in indian languages," in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[35] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., "Espnet: End-to-end speech processing toolkit," *Proc. Interspeech 2018*, pp. 2207–2211, 2018.

[36] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[38] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[39] Mahesh Viswanathan and Madhubalan Viswanathan, "Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech and Language*, vol. 19, no. 1, pp. 55–83, 2005.

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