An analysis of prosody modifications for the fusion of synthetic and natural data in low-resource ASR

Motivation

Q. How can we effectively use synthetic data along with natural speech to train ASR under low resource?

Possible Ways:

1. **Generate synthetic data more naturally (Synthetic close to natural) ✕ Resource constraint**
2. **Remove natural variations in natural data (Natural close to synthetic) ✕ Not sure how to**
3. **Modify both natural and synthetic (To make natural and synthetic close after modifications)**
    1. **Constant Pitch (Making pitch constant throughout the utterance i.e., Monotonous)**
    2. **Constant speaking rate (Duration)**
    3. **Constant energy**

Objectives

1.Selection of dataset

1. Audio: **Microsoft Telugu corpus** (**44882** audio clips **~40 Hrs**)
2. Text: **Vakyansh-LM Pre-processed text** from IndiCorp(AI4BHARATH)(#Sents: **35.7M**, #Words: **2L**)
    1.Selected **2 Lakh words**, generated pronunciations and trained 3-Gram LMs (Original and pruned)

2.Synthetic data generation

1. Synthesized **35173** audios using **IndicTTS (M&F)**

3.Pitch modification algorithm (Constant pitch throughout utterance) using PyWorld

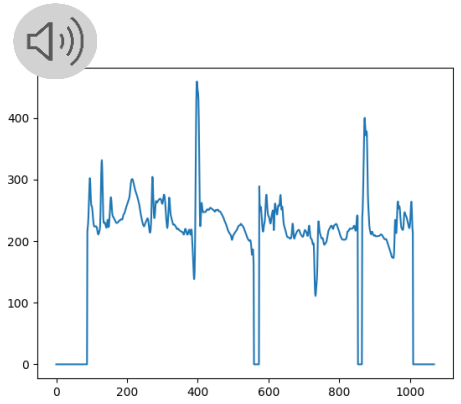
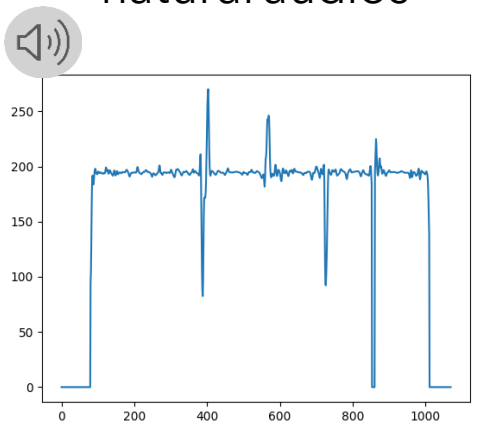
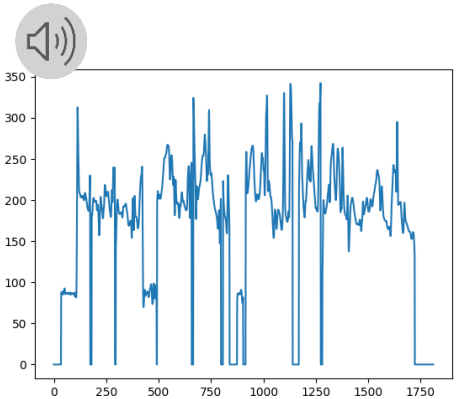
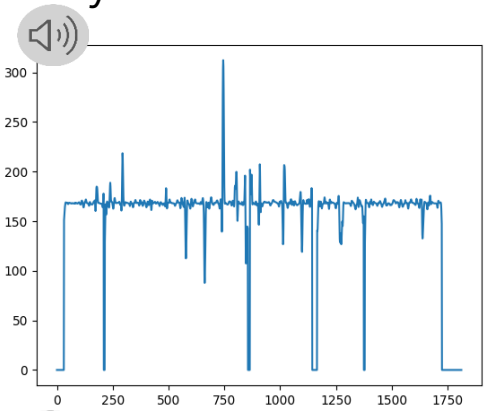
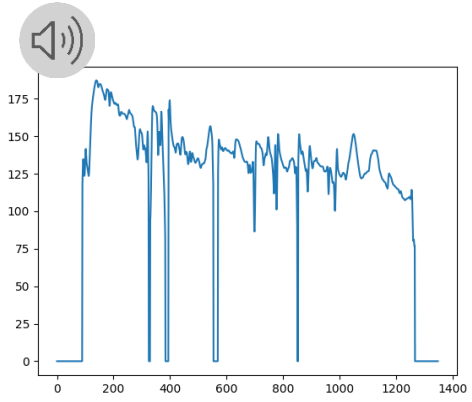
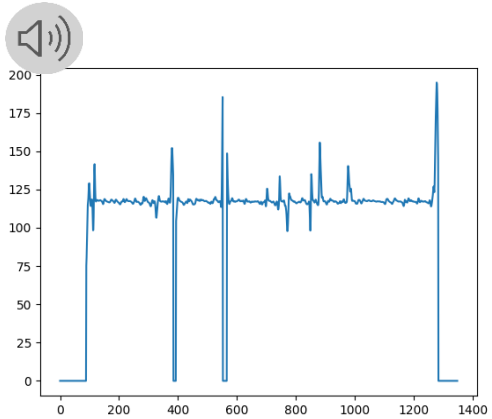
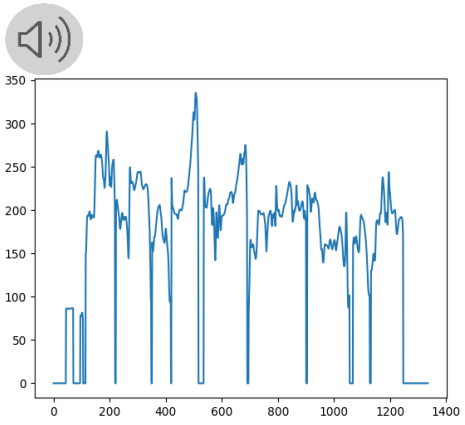
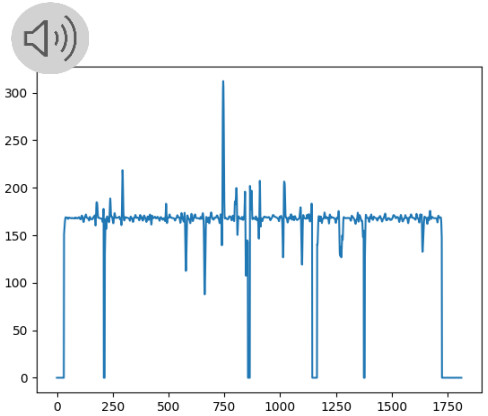
4.Training ASR with and without modification

1.Natural , 2. Natural + Synthetic and 3.Modified version of Natural + Synthetic

5.Testing

1.Natural, 2. Synthetic and 3. Modified version of Natural and Synthetic

Example audios (Natural, Synthetic and modified versions)

	Natural	Pitch modification on natural audios	Synthetic	Pitch modification on synthetic audios
Female				
Male				

Results

Table: Word Error Rate (WER) in percentage (%) of Natural and Synthetic audios tested using different ASR

ASR Model	Natural Test set	Synthetic Test set
ASR_Base	**18.40**	58.95
ASR_Combined	25.74	**14.04**
ASR _Combined_Modified	26.26	14.32

Observations:

1. Adding synthetic data reduced performance on natural data
2. No impact of Pitch Modification
    1. Not much pitch variation is seen (May be read speech)
    2. Quality of pitch modification method
3. Combined ASR performed well on synthetic data than natural
    1. Less number of speakers in synthetic (ASR became biased towards those speakers)
    2. Quality of synthetic speech

Result Contd..(Impact of extra large text)

Table: Word Error Rate (WER) in percentage (%) of Natural and Synthetic audios tested using different ASR

	Natural Test set		Synthetic Test set	
ASR Model	**LM Base**	**LM Extra Text**	**LM Base**	**LM Extra Text**
ASR_Base	18.40	**23.29**	58.95	**26.72**
ASR_Combined	25.74	**21.54**	14.04	**8.98**
ASR _Combined_Modified	26.26	**21.60**	14.32	**9.53**

Observation:

1. Addition of text improved WER on all cases except when natural test set tested on base ASR