Open Source Datasets: TTS, Machine Translation, Parallel corpora

B M Kalpajeet

Supervised by: Dr. Deepak KT
Department of Computer Science and Engineering
Indian Institute of Information Technology, Dharwad

19bcs117@iiitdwd.ac.in

January 15, 2022

Overview

1. Indic TTS and parallel data corpora
   - Datasets, Baseline Results
2. Other common datasets

IndicSpeech: Text-to-Speech Corpus for Indian Languages

- A 24 hour text-to-speech corpus for 3 major Indian languages: Hindi, Malayalam and Bengali.
- Aim of this paper was to verify the suitability of the IndicSpeech corpus for state-of-the-art TTS systems.
- Experiments: Deep Voice 3 was trained with the corpus (the trained model has since been released publicly.) A test set of 100 sentences was created.

IndicSpeech: baseline results

Error Type	Hindi	Malayalam	Bengali
Repeated words	4	10	6
Mispronunciations	11	18	14
Skipped words	1	6	3

Table 3: Scores for objective human evaluations. The participants were asked to identify the number of repeated words, mispronunciations and skipped words. One or more repeats, mispronunciations and skips count as a single mistake per utterance.

Model	Hindi	Malayalam	Bengali
Public API	2.98	3.32	3.63
Ours	**4.31**	**3.87**	**3.96**

Table 4: Average Mean-Opinion-Scores (MOS) for both Hindi, Malayalam and Bengali. The participants rate each of the speech outputs from both the models for each of the languages. The minimum score that can be given to an output is 1 and the maximum is 5. In our experiment, higher the score better the output from a model in terms on naturalness. (Murthy and others, 2016) has been used for the publicly available API.

Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages

- Contains a total of 49.7 million sentence pairs between English and 11 Indic languages (from two language families)
- Uses Samanantar to train a multilingual model named IndicTrans
- Observations: Sentence pairs included in Samanantar have high semantic similarity (found using Semantic Textual Similarity score, mean STS score of 4.27/5)
- Evaluation metric: BLEU scores (and some other modifications of it like sacreBLEU)

Samanantar: Results

- IndicTrans trained on Samanantar outperforms all publicly available open source models
- The absolute gain in BLEU score is higher for the Indic-En direction as compared to the En-Indic direction
- Performance gains are higher for low resource languages, especially Kannada and Odia, in the Indic-English direction

Samanantar: BLEU scores

	x-en									en-x								
Model	GOOG	MSFT	CVIT	OPUS	mBART	TF	mT5	IT	Δ	GOOG	MSFT	CVIT	OPUS	mBART	TF	mT5	IT	Δ
WAT2021																		
bn	20.6	21.8	-	11.4	4.7	24.2	24.8	**29.6**	4.8	7.3	11.4	12.2	-	0.5	13.3	13.6	**15.3**	1.7
gu	32.9	34.5	-	-	6.0	33.1	34.6	**40.3**	5.7	16.1	22.4	22.4	-	0.7	21.9	24.8	**25.6**	0.8
hi	36.7	38.0	-	13.3	33.1	38.8	39.2	**43.9**	4.7	32.8	34.3	34.3	11.4	27.7	35.9	36.0	**38.6**	2.6
kn	24.6	23.4	-	-	-	23.5	27.8	**36.4**	8.6	12.9	16.1	-	-	-	12.1	17.3	**19.1**	1.8
ml	27.2	27.4	-	5.7	19.1	26.3	26.8	**34.6**	7.3	10.6	7.6	11.4	1.5	1.6	11.2	7.2	**14.7**	3.3
mr	26.1	27.7	-	0.4	11.7	26.7	27.6	**33.5**	5.9	12.6	15.7	16.5	0.1	1.1	16.3	17.7	**20.1**	2.4
or	23.7	27.4	-	-	-	23.7	-	**34.4**	7.0	10.4	14.6	16.3	-	-	14.8	-	**18.9**	2.6
pa	35.9	35.9	-	8.6	-	36.0	37.1	**43.2**	6.1	22	28.1	-	-	-	29.8	31.	**33.1**	2.1
ta	23.5	24.8	-	-	26.8	28.4	27.8	**33.2**	4.8	9.0	11.8	11.6	-	11.1	12.5	13.2	**13.5**	0.3
te	25.9	25.4	-	-	4.3	26.8	28.5	**36.2**	7.7	7.6	8.5	8.0	-	0.6	12.4	7.5	**14.1**	1.7
WAT2020																		
bn	17.0	17.2	18.1	9.0	6.2	16.3	16.4	**20.0**	1.9	6.6	8.3	8.5	-	0.9	8.7	9.3	**11.4**	2.1
gu	21.0	22.0	23.4	-	3.0	16.6	18.9	**24.1**	0.7	10.8	12.8	12.4	-	0.5	9.7	11.8	**15.3**	2.5
hi	22.6	21.3	23.0	8.6	19.0	21.7	21.5	**23.6**	0.6	16.1	15.6	16.0	6.7	13.4	17.4	17.3	**20.0**	2.6
ml	17.3	16.5	18.9	5.8	13.5	14.4	15.4	**20.4**	1.5	5.6	5.5	5.3	1.1	1.5	5.2	3.6	**7.2**	1.6
mr	18.1	18.6	19.5	0.5	9.2	15.3	16.8	**20.4**	0.9	8.7	10.1	9.6	0.2	1.0	9.8	10.9	**12.7**	1.8
ta	14.6	15.4	17.1	-	16.1	15.3	14.9	**18.3**	1.3	4.5	5.4	4.6	-	5.5	5.0	5.2	**6.2**	0.7
te	15.6	15.1	13.7	-	5.1	12.1	14.2	**18.5**	2.9	5.5	7.0	5.6	-	1.1	5.0	5.4	**7.6**	0.7
WMT																		
hi	31.3	30.1	24.6	13.1	25.7	25.3	26.0	**29.7**	-1.6	24.6	24.2	20.2	7.9	18.3	23.	23.8	**25.5**	0.9
gu	30.4	29.9	24.2	-	5.6	16.8	21.9	**25.1**	-5.4	15.2	17.5	12.6	-	0.5	9.0	12.3	**17.2**	-0.3
ta	27.5	27.4	17.1	-	20.7	16.6	17.5	**24.1**	-3.4	9.6	10.0	4.8	-	6.3	5.8	7.1	**9.9**	-0.1
UFAL																		
ta	25.1	25.5	19.9	-	24.7	26.3	25.6	**30.2**	3.9	7.7	10.1	7.2	-	9.2	11.3	**11.9**	10.9	-1.0
PMI																		
as	-	16.7	-	-	-	7.4	-	**29.9**	13.2	-	10.8	-	-	-	3.5	-	**11.6**	0.8

PMIndia: A Collection of Parallel Corpora of Languages of India

- Consists of parallel sentences which pair 13 major languages of India with English, up to 56000 sentences for each language pair.

Language	en-X	X-en
as	5.3	8.5
bn	6.6	10.9
gu	8.8	16.1
hi	23.4	19.6
kn	7.7	14.9
ml	1.8	10.1
mni	11.5	13.9
mr	6.0	12.4
or	9.2	12.4
pa	18.1	18.8
ta	3.2	11.4
te	7.4	14.3
ur	16.5	15.3

Table 5: Bleu scores for NMT systems built on the PMIndia corpus, for English to/from languages of India, setting aside 1000 sentences for dev and 1000 for test.

IndicTTS (IIT Madras)

- Contains over 10,000 mono and English spoken sentences/utterances recorded by both male and female native speakers, in 13 major languages.
- The voice building module has 5 components: a common label set, parsing and unified parser, hybrid segmentation, pruning and HTS.
- There are also android applications using this TTS system: Indic TTS, Tamil TTS and Hindi TTS

The IIIT-H Indic Speech Databases (2012)

- This speech database contains text and speech data in 7 languages.
- For each language, 1000 sentences have been selected, derived from Wikipedia articles. The corpus covers roughly about 5000 most frequent words in the corresponding languages.
- Mel-cepstral distortion (MCD) scores were calculated for two different voice types.

Crowdsourced high-quality Kannada multi-speaker speech dataset

- These are about 4400 sentences collected from native Kannada speakers who voluntarily supplied the data.
- This mono audio is of high quality at 48kHz, 16 bit and recorded in a quiet environment.

The IIT Bombay English-Hindi Parallel Corpus

- The corpus contains 1.6 million parallel segments, which have been pre-processed for machine translation
- The training corpus includes utterances, phrases, and dictionary entries from a variety of applications and domains, such as from government websites, legal judgements, recognized dictionaries, etc

System	eng-hin		hin-eng	
	BLEU	*METEOR*	*BLEU*	*METEOR*
SMT	11.75	0.313	14.49	0.266
NMT	12.23	0.308	12.83	0.219

Table 3: Results for Baseline Systems

Multiple Language datasets (Indian Language Technology Proliferation Deployment Centre)

- Language data from 22 Indian languages is included in this dataset, each having about 30000-40000 audio files.
- This was created primarily to test the support for Indian languages on mobile devices.
- Mainly comprises of agriculture related speech data.

LibriTTS

- Multi-speaker English corpus with about 585 hours of read English speech at a sampling rate of 24kHz
- The LibriTTS dataset differs from the LibriSpeech dataset in the sense that there are sentence breaks at which the speech is split and contextual information in the text can be extracted.
- Experimental results reveal that in five out of six evaluation speakers, neural end-to-end TTS models trained from the LibriTTS corpus scored above 4.0 in mean opinion scores in naturalness.

The LJ Speech Dataset

- English speech dataset which contains about 13000 audio clips (sample rate of 22 kHz) of a single speaker uttering passages from non-fiction books
- Used in the text-to-speech synthesis task in the paper 'Neural Speech Synthesis with Transformer Network'
- Proposed model achieves state-of-the-art performance (outperforms Tacotron2 with a gap of 0.048) and is very close to human quality (4.39 vs 4.44 in MOS)
- For efficiency, the Transformer TTS network can speed up the training about 4.25 times faster compared with Tacotron2.

Updated dataset link

The following document will be regularly updated with further open source datasets for TTS, ASR, machine translation, parallel corpora:

- links-open-source-dataset

Resources: datasets (links)

- Language resources by Google
- IITG Multivariability Speaker Recognition Database
- IndicNLP AI4Bharat: A Catalog of resources for Indian language NLP
- IIT Kharagpur SHRUTI Bengali Continuous ASR Speech Corpus
- Microsoft Speech Corpus (Telugu, Tamil and Gujarati)
- Universal Language Contribution API

Resources: models and NLP (links)

- Indic NLP library
- Neural TTS for Kannada
- Odia NLP
- Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation
- ESPnet: Multilingual end-to-end speech translation
- Mozilla TTS
- CoquiAI TTS
- YourTTS: Towards Zero-Shot Multi-Speaker TTS

References

- Srivastava, Mukhopadhyay, Prajwal, Jawahar (IIIT-H) "IndicSpeech: Text-to-Speech Corpus for Indian Languages"
- Ramesh et al. (IIT M, Microsoft Research, EkStep, AI4Bharat) "Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages"
- Haddow, Kirefu. "PMIndia- A Collection of Parallel Corpora of Languages of India"
- Prahallad et al. "The IIIT-H Indic Speech Databases"
- Baby, Thomas, Nishanthi, TTS Consortium. "Resources for Indian languages"

References

- Kunchukuttan, Mehta, Bhattacharyya "The IIT Bombay English-Hindi Parallel Corpus"
- Ito and Johnson. "The LJ Speech Dataset"
- Zen et al. "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech"
- https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html
- https://paperswithcode.com/task/semantic-textual-similarity
- https://ai.facebook.com/blog/the-flores-101-data-set-helping-build-better-translation-systems-around-the-world/

Thank You