Demo Paper Project Page

MacST: Multi-Accent Speech Synthesis via Text Transliteration for Accent Conversion

Authors: Sho Inoue, Shuai Wang, Wanxing Wang,
Pengcheng Zhu, Mengxiao Bi, Haizhou Li

Presenter: Sho Inoue

Email: shoinoue@link.cuhk.edu.cn

网易伏羲

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

深圳市大数据研究院
Shenzhen Research Institute of Big Data

Introduction

- **Foreign Accent Conversion**: Convert the accent of the source speech
  - While keeping the linguistic content and the speaker identity.

- **Problem**: Lack of parallel dataset with only accent changes.
- **Solution**: To generate the target samples to build synthetic parallel dataset.
  - However, it can lead some problems.
    - Entanglement issue of speaker and accent
    - Limited availability of accented speeches.

⟶ We propose a pipeline to address these issues using **text transliteration**.

Proposed System

- MacST: Multi-accent speech synthesis via text transliteration to construct parallel accent dataset
  - **Translation**: Converting the language while keeping <u>the similar meaning</u>.
  - **Transliteration**: Converting the language while keeping <u>the phonetic similarity</u>.
- Procedure:
  - Describe English sentences using the characters of the target language (**transliteration**).
    - E.g. *I love you* → <u>*आई लव यू (aaee lav yoo)*</u> or <u>アイラブユー (*ai rabu yū*)</u>
  - Use **Multilingual TTS** to generate the accented speech from transliterated texts.

American Hindi Korean Japanese

"Again he had done the big thing."

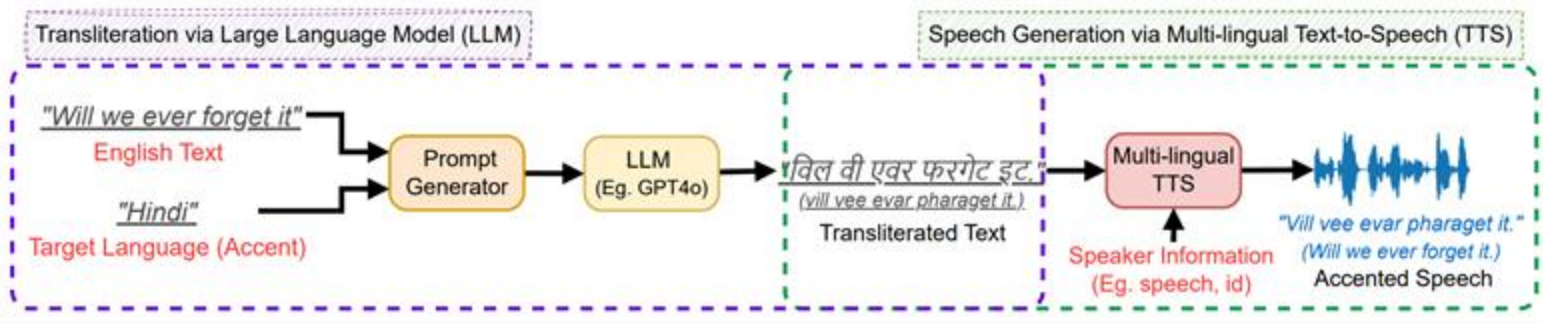
Hypothesis and Motivation

- **Hypothesis**:
  - Lexical features of accents are based on availability of phonemes in the native languages [1-2].
  - Large Language Models (LLMs) is capable of transliterating texts of target languages.

- **Motivation**:
  - We do not need accented speech samples.
  - We can avoid entanglement of speaker and accent.
  - MacST applies to *any* English texts and *any* speaker.
  - Consistent linguistic representation in accented speech across different speakers.

[1] Alison Behrman: Segmental and prosodic approaches to accent management
[2] James Flege: Second language speech learning: Theory, findings and problems

Methodology



- Two Procedures:
  - Text Transliteration via Large Language Models (LLMs)
  - Speech Generation via Multilingual Text-to-Speech (TTS) Models
- Three inputs: English Text, Target Accent, and Speaker Information
- Speaker Information depends on TTS models.
  - In this paper, speech samples from the chosen speaker.
- This system can be applied to various LLMs and Multilingual TTS models.

Dataset Analysis
(Experiment Setup)

- **Target Accents**: American, Hindi, and Korean
- We synthesize accented speech samples using native and non-native speakers.
  - Native Speaker: Accent Addition
  - Non-native Speaker: Accent Enhancement
- **Evaluation Metrics**: MUSHRA tests for Speech Naturalness (Humanness) and Accentedness
- **Comparing Datasets**: L2-ARCTIC and CMU-ARCTIC
  - ARCTIC datasets contain speech samples from different speakers with the same transcripts.
    - *Each speaker only speaks in a single accent*.

Dataset Analysis (Results)

- **American Speakers**: SLT
- **Hindi Speakers**: ASI, TNI
- **Korean Speakers**: HKK, YDCK
- **MacST**: Proposed system
- The language in MacST indicates the transliteration language.
- Accent Addition Capability is good.
  - E.g. SLT/American → SLT/Hindi
- Accent Enhancement Capability is also good.
  - Ground-Truth → MacST for ASI/Hindi, TNI/Hindi, HKK/Korean, and YDCK/Korean

	Naturalness (↑)	Accentedness (↑)
Ground-Truth (SLT/American)	76.48± 3.82	9.56± 1.32
MacST (SLT/*American*)	70.95± 4.07	10.78± 1.41
Ground-Truth (ASI/Hindi)	85.17± 1.87	67.67± 2.60
Ground-Truth (TNI/Hindi)	81.29± 2.76	70.74± 2.40
MacST (SLT/*Hindi*)	69.51± 3.99	51.61± 3.02
MacST (ASI/*Hindi*)	82.12± 2.36	73.61± 2.51
MacST (TNI/*Hindi*)	79.64± 2.82	77.35± 2.66
Ground-Truth (SLT/American)	66.84± 3.45	6.90± 1.07
MacST (SLT/*American*)	70.37± 3.52	8.56± 1.40
Ground-Truth (HKK/Korean)	75.28± 2.55	39.08± 2.46
Ground-Truth (YDCK/Korean)	78.84± 1.87	32.90± 2.10
MacST (SLT/*Korean*)	58.47± 4.85	77.63± 2.33
MacST (HKK/*Korean*)	63.22± 4.06	83.40± 1.67
MacST (YDCK/*Korean*)	63.87± 4.36	83.44± 1.67

Accent Conversion (Experiment Setup)

- **Accent Conversion**: American → Hindi
- We synthesize accented speech samples using American speaker (SLT).
- **Evaluation Metrics**:
  - MUSHRA tests for Speech Naturalness (Humanness) and Accentedness
  - Objective Evaluations:
    - Speech Intelligibility: Word Error Rate (WER)
    - Speaker Similarity: Speaker Encoding Cosine Similarity (SECS)
    - Accentedness: Accent classification prob (Hindi)
- **Two accent conversion models with different training datasets**:
  - The parallel dataset with the ground-truth source and the synthetic target (1 hour pairs)
  - Additional pairs of the synthetic source and the target (additional 4 hours): **Data Augmentation**

Accent Conversion (AC)
(Results)

	Speech Quality		Accentedness			Speaker Similarity
	MUSHRA (↑)	WER (↓)	MUSHRA (↑)	Classification Prob. (↑)	AECS Diff. (↑)	SECS (↑)
Ground-Truth (American)	76.48 ± 3.82	1.97	9.56 ± 1.32	0.000	-	-
MacST (American)	70.95 ± 4.07	1.75	10.78 ± 1.41	0.000	-	0.866
MacST (Hindi)	69.51 ± 3.99	8.52	51.61 ± 3.02	0.819	-	0.822
AC w/o Data Augmentation	51.48 ± 3.73	13.99	34.85 ± 2.29	0.801	0.411	**0.834**
AC w/ Data Augmentation (ours)	**67.18** ± 3.43	**8.74**	**47.26** ± 2.65	**0.897**	**0.465**	0.833

- Consistent speaker characteristics between the source and the converted audio.
- Accent conversion significantly increased accentedness: Ground-Truth (American) → AC results
- Data Augmentation enhanced the conversion results in speech quality and accentedness.

Conclusion

- We introduce the multi-accent speech synthesis via text transliteration method (MacST)
  - Transliteration via LLMs
  - Speech Generation via Multilingual TTS Models

- Dataset analysis validates MacST's ability to amplify accents in native and non-native English speakers.

- Experiment results validate the efficacy of our method in training accent conversion models.

Demo Paper Project Page

Thank you very much for your time
Q & A

Presenter: Sho Inoue

Email: shoinoue@link.cuhk.edu.cn

Accent Conversion (Model Configuration)

- **Accent Conversion (AC)**: Voice Transformer Networks (VTN)
  - A sequence-to-sequence encoder-decoder model.
  - Mel-spectrogram as input and output.
  - We pretrain AC with TTS-like tasks using LibriTTS-R.
- **Pre-training Strategy**: Two-stage pretraining.
  - 1st stage: Input is Hubert discrete tokens (without repetition) and Output is Mel-spectrogram
  - 2nd stage: Input and Output are Mel-spectrograms
    - Initialize encoder and Freeze decoder
- **Vocoder**: HiFiGAN trained on LibriTTS-R and ARCTIC datasets.

Text Transliteration with LLMs

- **How to build a prompt for LLMs**
  - Word-level transliteration with phoneme-sequence.
  - Put three candidate words and Sort them in similarity order.
  - We include few transliterated samples to avoid *translation*.
- **Post Process**
  - We got six responses in total, three for GPT 3.5 Turbo and three for GPT-4o.
  - Among six responses, we obtain the most frequent transliterated texts for each word.
  - We re-put commas and periods.

Speech Generation with Multilingual TTS

- **Multilingual TTS Models**: the Eleven Multilingual v2 model from 11Elevenlabs.
    - It covers 29 languages.
    - Speaker Condition: speech samples (voice clone)
    - Language Condition: the characters of the input text

Evaluation Metrics (Accentedness)

- We used three metrics to evaluate "Accentedness" of synthesized speeches.
  - MUSHRA test for Accentedness (strength of the accent)
  - Classification probability for Hindi accent using a pre-trained accent classification model.
  - Accent Encoding Cosine Similarity (AECS) Difference.
- **AECS Difference**: To quantify accent similarity of converted speech from native and non-native speech.
  - Obtain accent embeddings of converted sample and MacST samples of American and Hindi speech.
  - Calculate AECS between
    - Converted speech and American speech: AECS_american
    - Converted speech and Hindi speech: AECS_hindi
  - compute "AECS_hindi - AECS_american"

Demo

Paper

Project Page

Thank you very much for your time
Q & A

Presenter: Sho Inoue

Email: shoinoue@link.cuhk.edu.cn