Paper ID: 1763

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


Sho Inoue [1,2,3], Shuai Wang [1,2], Wanxing Wang [3], Pengcheng Zhu [3], Mengxiao Bi [3], Haizhou Li [1,2]

[1] *School of Data Science* [2] *Shenzhen Research Institute of Big Data*
*The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, China*
[3] *Fuxi AI Lab, NetEase Inc., Hangzhou, China*

Email:shoinoue@link.cuhk.edu.cn

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

网易伏羲

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


2025 ICASSP HYDERABAD, INDIA Celebrating Signal Processing

Demo Paper

## Introduction

**Foreign Accent Voice Conversion:**
Convert the accent of the source speech while keeping the linguistic content and the speaker identity

**Major Problem:**
Lack of parallel dataset with only accent changes

**MacST's Goal:**
Multi-accent speech synthesis via text transliteration to construct parallel accent dataset

## Transliteration

**Translation:**
Converting the language while keeping the similar meaning.

**Transliteration:**
Converting the language while keeping the phonetic similarity.

| Language | Transliteration ("Accent") | Pronunciation |
|---|---|---|
| Hindi | अकसएन्ध | aksemt |
| Japanese | アクセント | akusento |
| Korean | 액센트 | aegsenteu |

## Motivation & Contributions

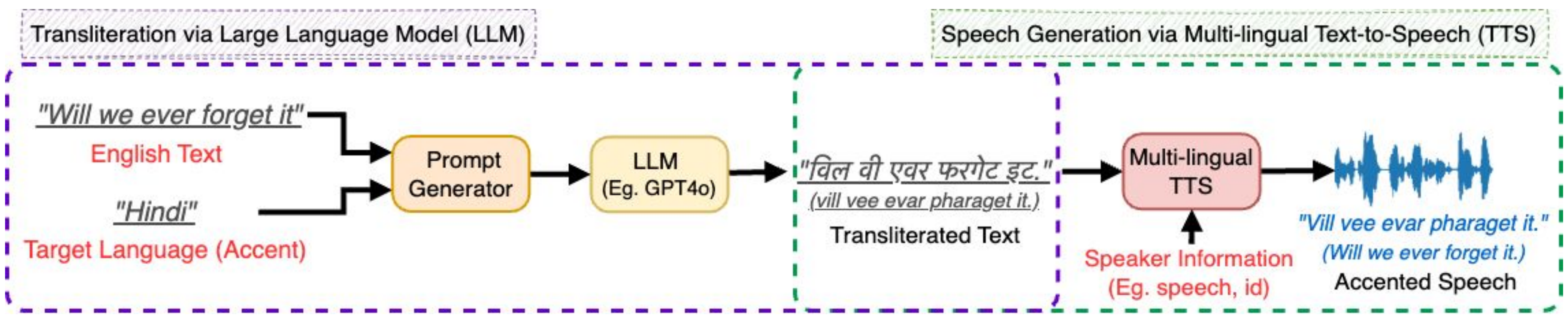
**Motivations:**
(1) No need accented speech samples
(2) No entanglement issue of speaker and accent.
(3) Applicable to *any* English texts and *any* speaker
(4) Consistent linguistic representation across different speakers.

**Contributions:**
(1) Pipeline for generating multi accent speech using pretrained LLMs and multilingual TTS models.
(2) First study of transliteration in building a parallel accent dataset and apply it to accent conversion.

Novel approach to synthesize accented speech that is applicable to various pretrained LLMs and Multilingual TTS!

## Conclusion

(1) Dataset analysis validates MacST's ability to enhance accents in native and non-native English speakers
(2) Subjective and objective metrics confirm effectiveness of MacST in training foreign accent conversion models.




## MacST Methodology

**Overall Diagram with Two Steps (Figure):**
**(1) Transliteration via LLMs (2) Speech Generation via Multilingual TTS**
Three Inputs: English Text, Target Accent, Speaker Information

**Text Transliteration via Large Language Models (LLMs):**
Prompt Engineering for LLMs:
(1) Word-level transliteration with the corresponding phoneme sequence.
(2) Three transliteration texts per word.
(3) Few-shot examples to avoid *translation.*
Get three responses from two pretrained LLMs: GPT-3.5 Turbo and GPT-4o

**Speech Generation via Multilingual Text-to-Speech (TTS):**
We used Multilingual model from 11Elevenlabs, covering 29 languages.

## Experiments and Results

**Dataset Analysis:**
We analyzed synthesized speech samples with the existing accent corpus.
**Target Accents:**
American, Hindi, Korean
**Evaluation Metrics:**
MUSHRA tests for speech naturalness (humanness) and accentedness.
**Comparing Datasets:**
L2-ARCTIC and CMU-ARCTIC

| | 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 |

**Language* in MacST indicates the transliteration language
*American Speaker Hindi Speaker Korean Speaker

**Results:**
(1) Accent Addition ability is Good👍 E.g. MacST: SLT/American → SLT/*Hindi*
(2) Accent Enhancement ability is Good👍 E.g. Ground-Truth → MacST for Hindi and Korean speakers

| | Speech Quality | | Accentedness | | |
|---|---|---|---|---|---|
| | MUSHRA (↑) | WER (↓) | MUSHRA (↑) | Classification Prob. (↑) | AECS Diff. (↑) |
| 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 | - |
| MacST (Hindi) | 69.51± 3.99 | 8.52 | 51.61± 3.02 | 0.819 | - |
| AC w/o Data Augmentation | 51.48± 3.73 | 13.99 | 34.85± 2.29 | 0.801 | 0.411 |
| AC w/ Data Augmentation (ours) | **67.18**± 3.43 | **8.74** | **47.26**± 2.65 | **0.897** | **0.465** |

**Accent Voice Conversion (AC):**
We built accent parallel dataset for American → Hindi accent conversion.
**Speaker**: SLT (female American speaker) from CMU-ARCTIC
**Subjective Evaluation Metrics**: MUSHRA tests
**Objective Evaluation Metrics**:
Word Error Rate (WER) and Accent Classification Probability (Hindi)
**Comparing two accent conversion models with different training data**:
(1) The parallel dataset with ground-truth source and synthetic target (1 hour pairs)
(2) Data Augmentation: Additional pairs of synthetic source and target (*additional* 4 hours).
**Results:**
(1) Accent conversion significantly increased accentedness:
*Ground-Truth (American) → AC*
(2) Data Augmentation enhanced speech quality and accentedness.