# Accentron: Foreign Accent Conversion to Arbitrary Non-Native Speakers using Zero-Shot Learning (Ding et al., 2022)

Group 3

# Before we start …

When British People Say Water In The USA

# Outline

1. Introduction
2. Methods
3. Experimental setup
4. Results
5. Limitations and future directions

# 1. Introduction

# 1.1 Motivation

Foreign accent conversion (FAC) seeks to generate a new voice that retains the voice identity of a second language (L2) speaker while incorporating the accent of a native (L1) speaker.

Applications:

1. Pronunciation training
2. Movie dubbing
3. Personalized text-to-speech (TTS) synthesis
4. ...

# 1.2 Previous Research

Related work:

- Foreign accent conversion:
    - Building an articulatory synthesizer for the L2 speaker
    - Acoustic methods
        - Framer pairing methods
        - Sequence to sequence (seq2seq )methods
- Many-to-many voice conversion (VC)
    - Traditional VC approaches
    - Many-to-many VC approaches

# 1.2 Previous Research

Limitations:

- They require training a separate model for each pair of L1 and L2 speakers.
- They demand a substantial quantity of speech data, approximately 1000 utterances, for each L2 speaker.

# 1.3 Current Study

The present research:

- introduces a zero-shot learning method of FAC capable of synthesizing speech for L2 speakers not included in the training data.
- doesn’t require large quantity of speech data (a few second of audio) from the L2 speaker not in the training data.

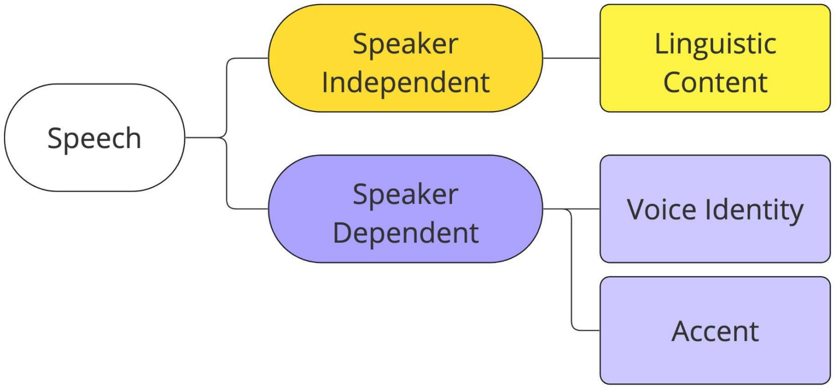
# 1.4 Demo

https://shaojinding.github.io/samples/accentron/

# 2. Methods

# 2. Methods: Architecture

What we need?



# 2. Methods: Architecture

- Parallel Training! (note that source and target speech are all the same sentence but different speaker)

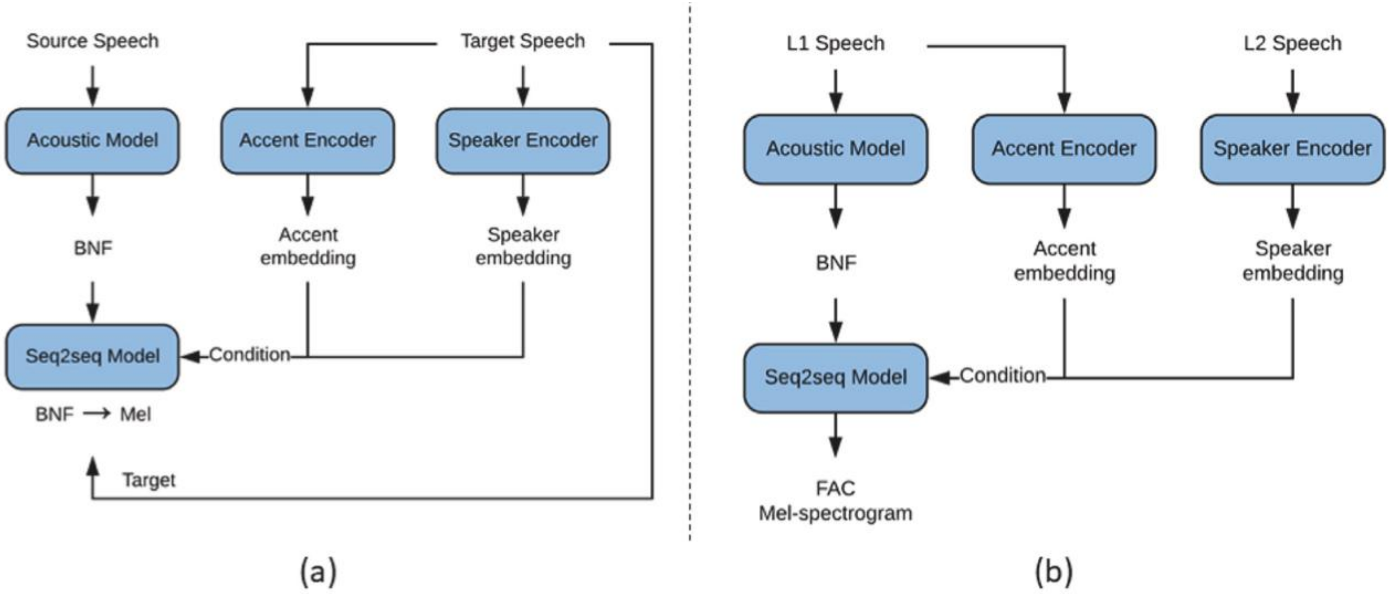


**Fig. 1.** (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

# 2. Methods: Architecture

- Parallel Training! (note that source and target speech are all the same sentence but different speaker)

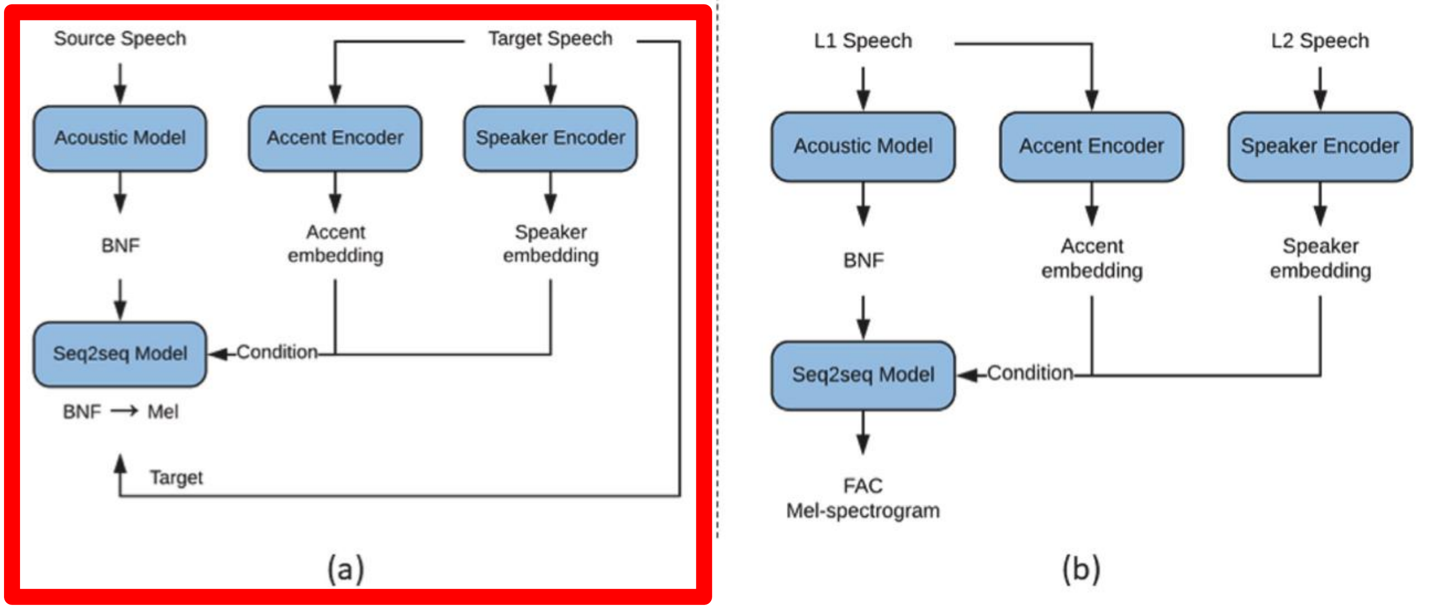


**Fig. 1.** (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

# 2. Methods: Architecture

- Parallel Training! (note that source and target speech are all the same sentence but different speaker)

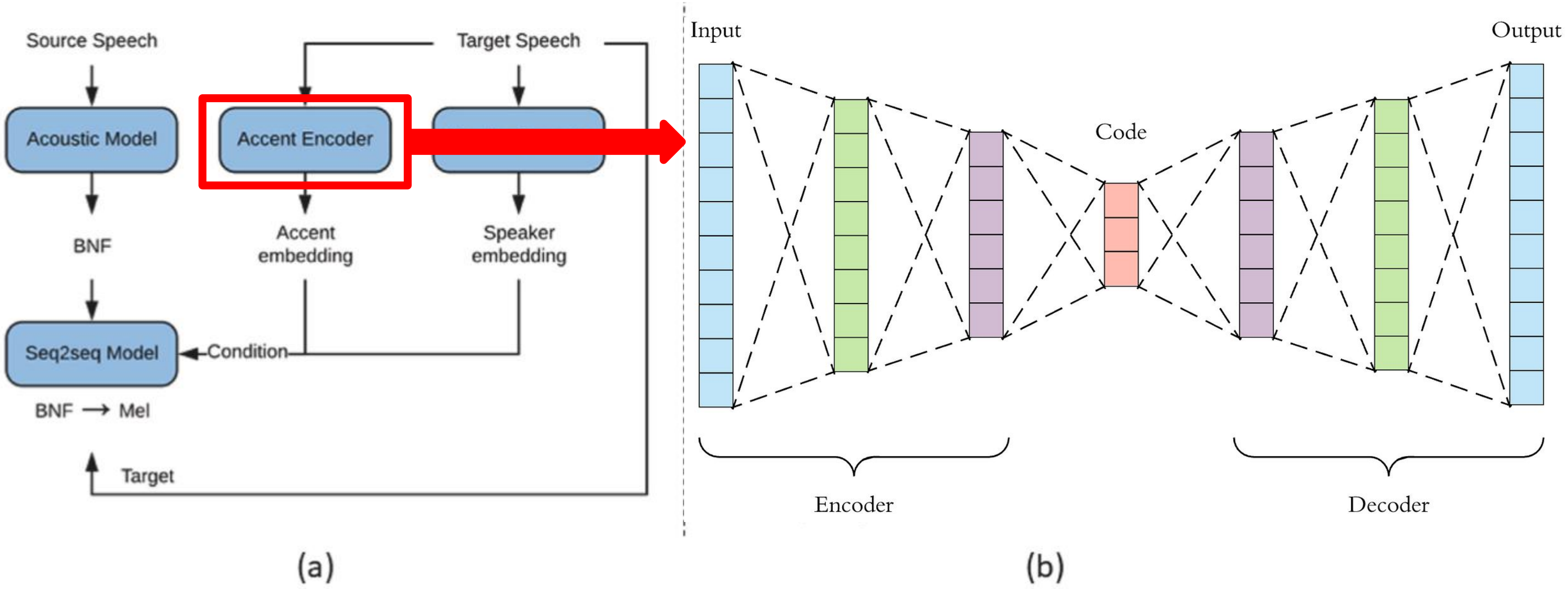


**Fig. 1.** (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

# 2. Methods: Architecture

- Parallel Training! (note that source and target speech are all the same sentence but different speaker)

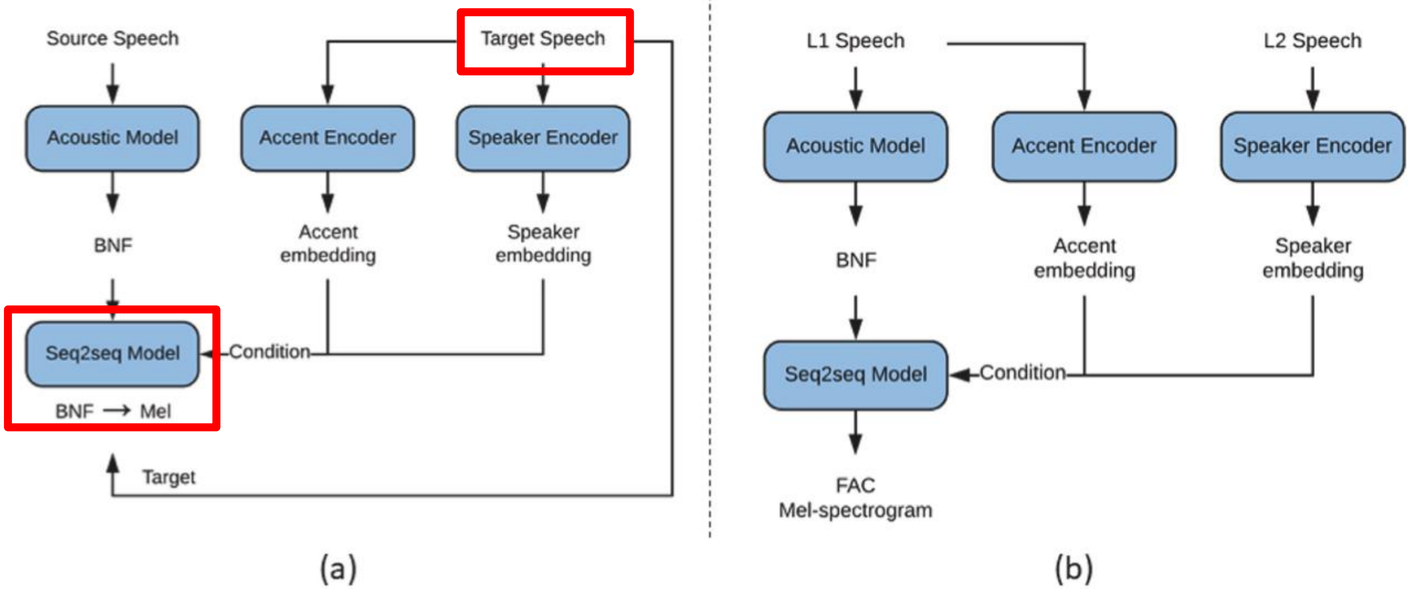


**Fig. 1.** (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

# 2. Methods: Architecture

- Parallel Training! (note that source and target speech are all the same sentence but different speaker)

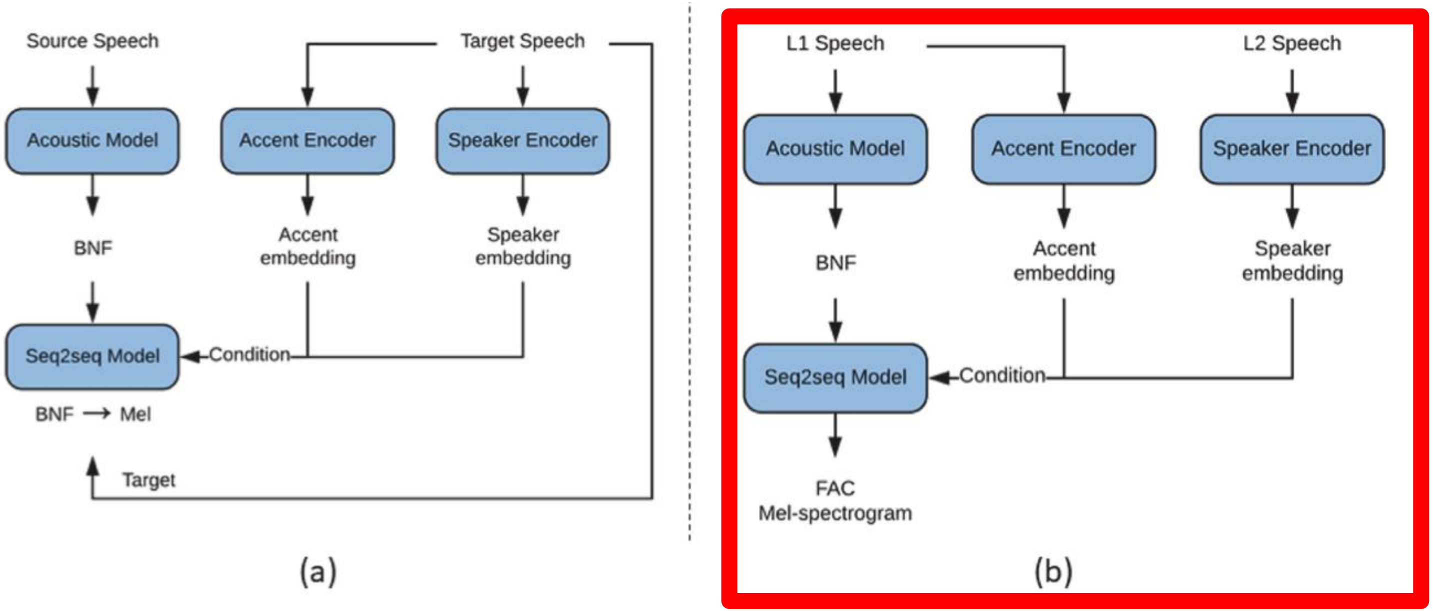


**Fig. 1.** (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

# 2. Methods: Architecture

- Parallel Training! (note that source and target speech are all the same sentence but different speaker)

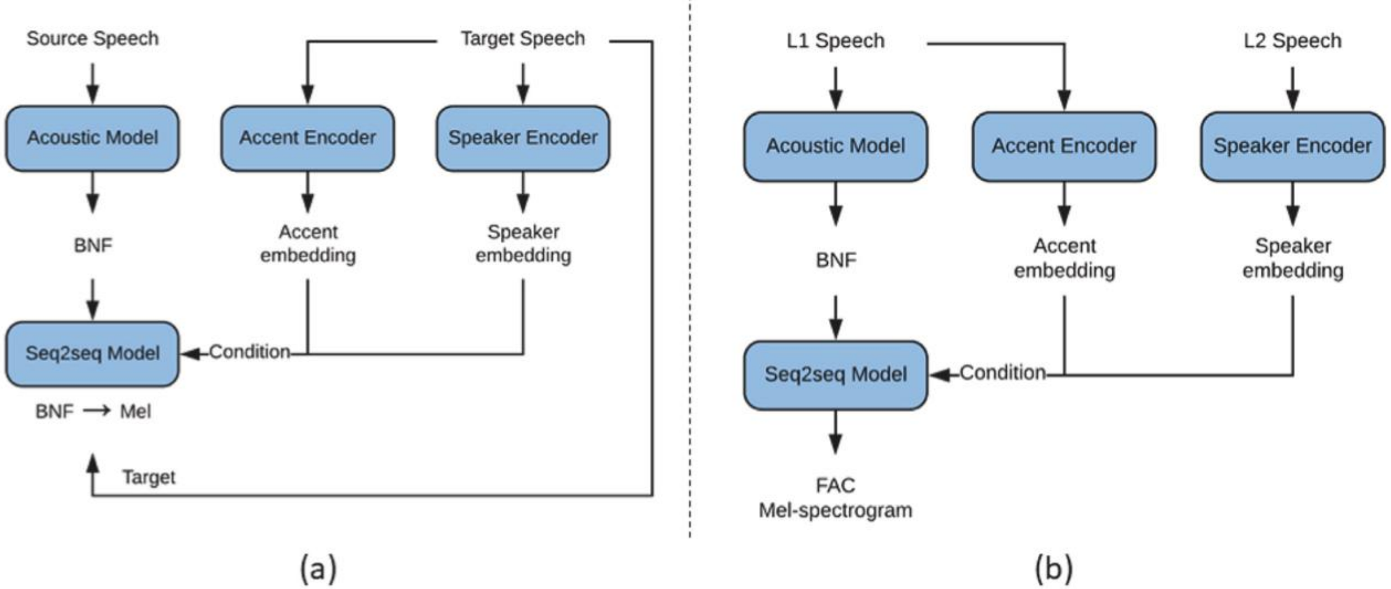


**Fig. 1.** (a): Overall training workflow of Accentron. (b): Overall inference workflow of Accentron. Source: a selected reference L1 speaker, Target: any L1/L2 speaker, BNF: bottleneck feature, L1: native, L2: non-native. Each of the modules is trained independently.

# 2. Methods: Acoustic Model

Features from of a pre-trained phoneme recognition model

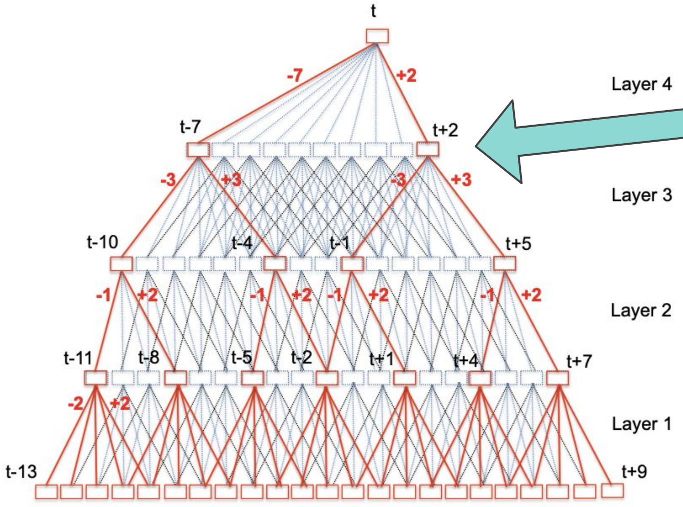


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

Last layer is extracted as BNF (bottleneck feature)

# 2. Methods: Speaker/Accent Encoders

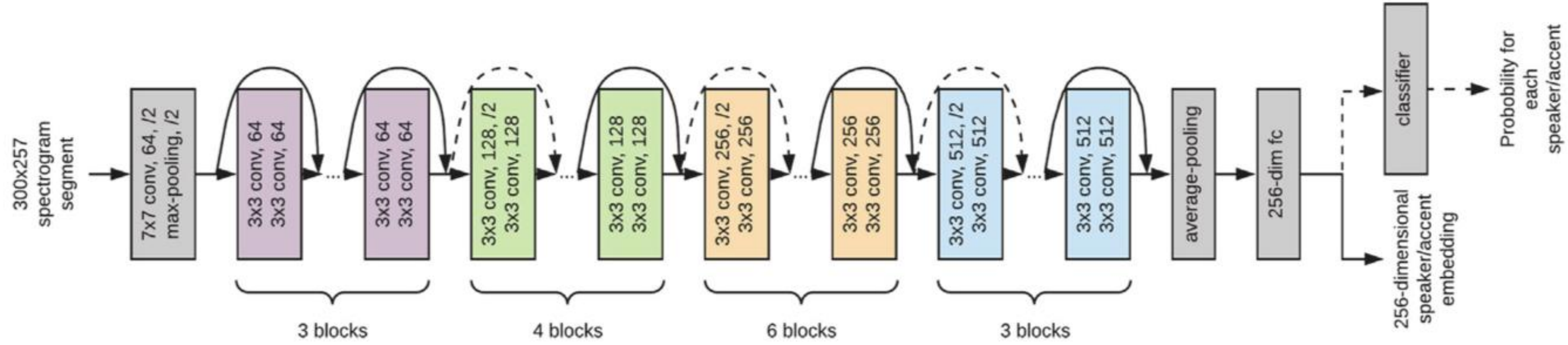

300x257 spectrogram segment
7x7 conv, 64, /2 max-pooling, /2
3x3 conv, 64 3x3 conv, 64
3x3 conv, 64 3x3 conv, 64
3x3 conv, 128, /2 3x3 conv, 128
3x3 conv, 128 3x3 conv, 128
3x3 conv, 256, /2 3x3 conv, 256
3x3 conv, 256 3x3 conv, 256
3x3 conv, 512, /2 3x3 conv, 512
3x3 conv, 512 3x3 conv, 512
average-pooling
256-dim fc
classifier
Probability for each speaker/accent
256-dimensional speaker/accent embedding
3 blocks
4 blocks
6 blocks
3 blocks

**Fig. 2.** Speaker/accent encoder model architecture. The model is based on ResNet-34 (He et al., 2016). Each convolution block is illustrated as the kernel size and channel numbers. "/2" means the layer divides the spatial resolution by 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

# 2. Methods: Seq2seq foreign accent conversion model

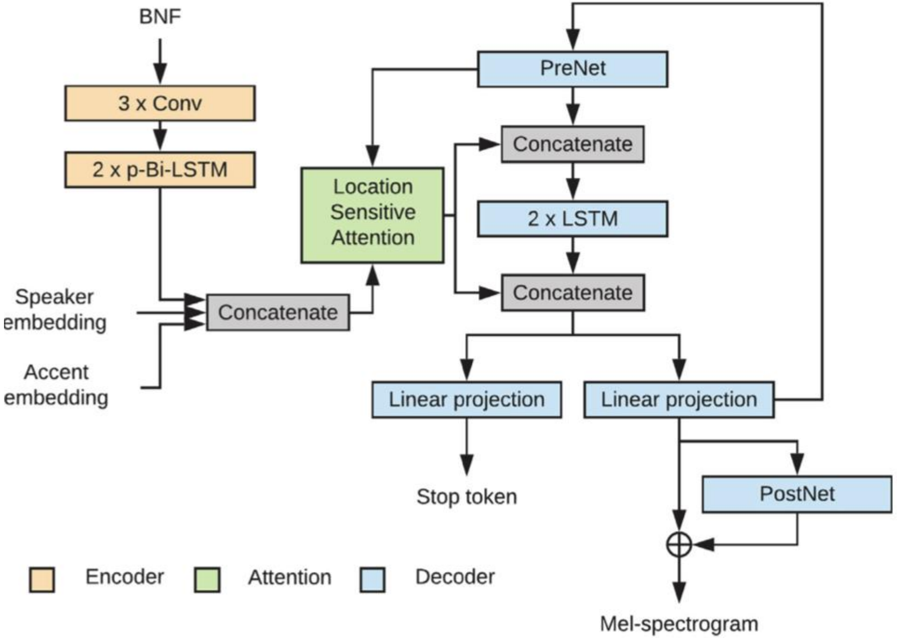


**Fig. 3.** The seq2seq model in Accentron.

# 3. Experimental setup

# 3. Experimental Setup

- Acoustic Model
  - Librispeech
- Speaker Encoder
  - VoxCeleb1 corpus
- Accent Encoder
  - Speech Accent Archive
    - “Please call Stella” paragraph
  - Subset of accents that had at least 30 speakers
  - 18 accents
- Seq2seq foreign accent conversion model evaluation
  - ARCTIC and L2-ARCTIC corpora

# 4. Results

# 4.1 Objective Evaluation

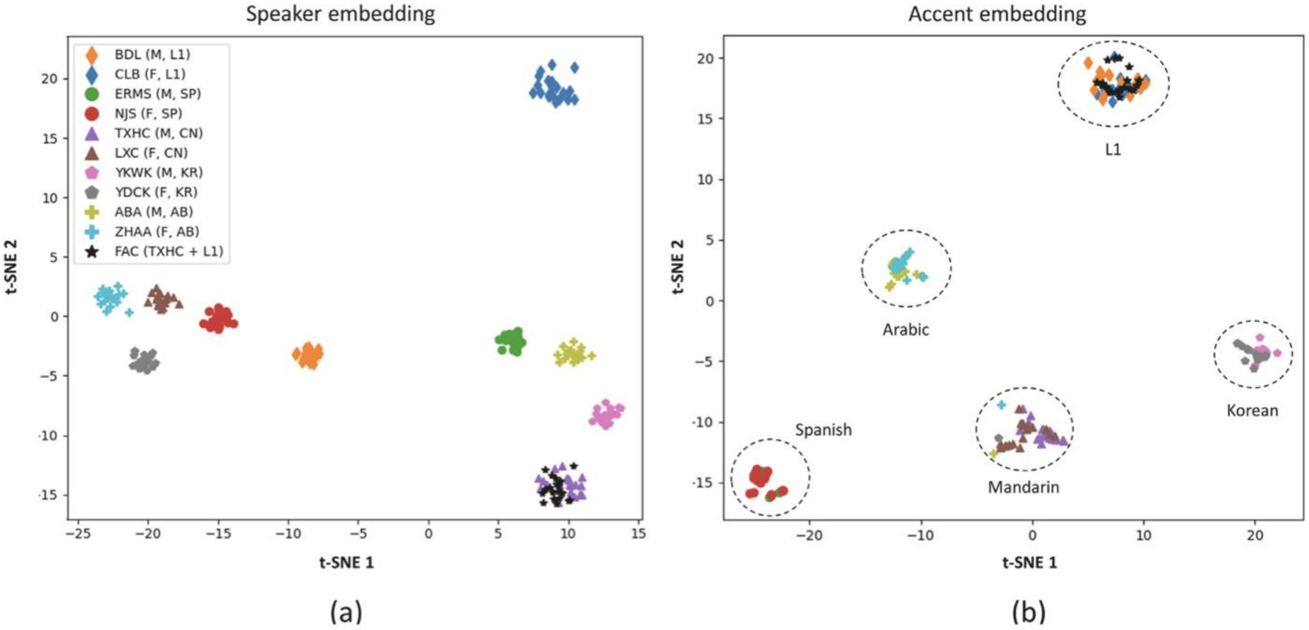


**Fig. 4.** Speaker and accent embedding visualization of FAC syntheses for TXHC using t-SNE. (a): speaker embedding; (b): accent embedding. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

# 4.1 Objective Evaluation

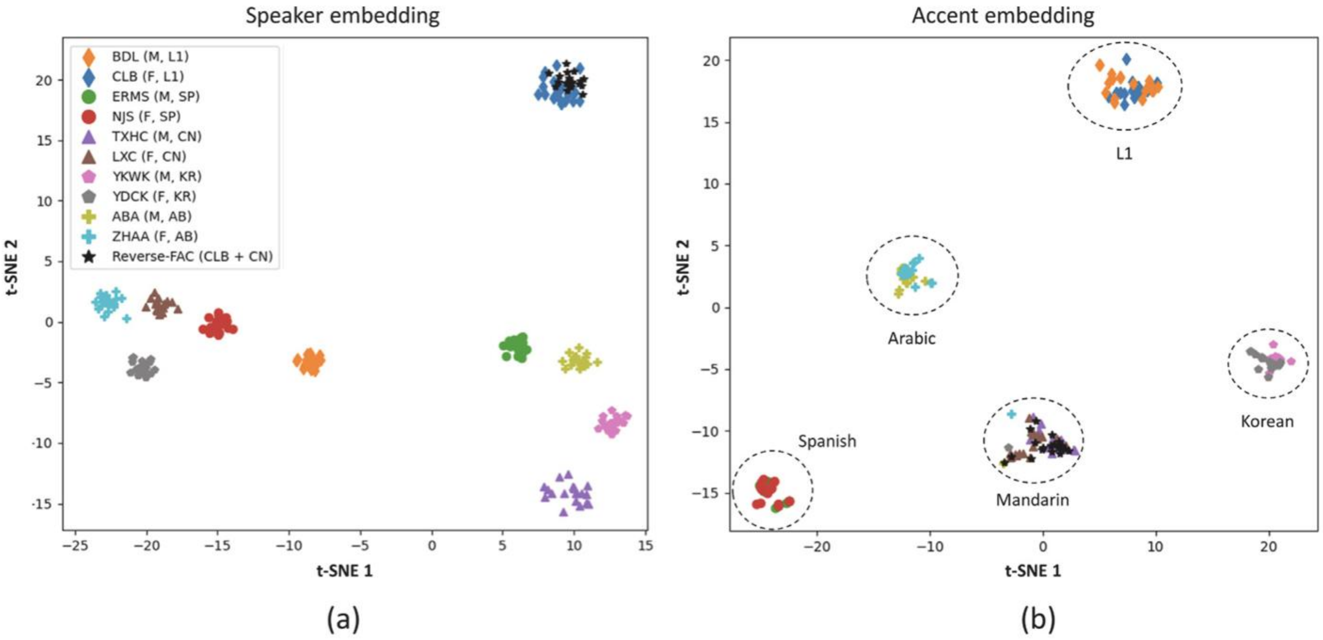


**Fig. 5.** Speaker and accent embedding visualization of *reverse* FAC syntheses (CLB with a Mandarin accent) using t-SNE. (a): speaker embedding; (b): accent embedding. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

# 4.1 Objective Evaluation

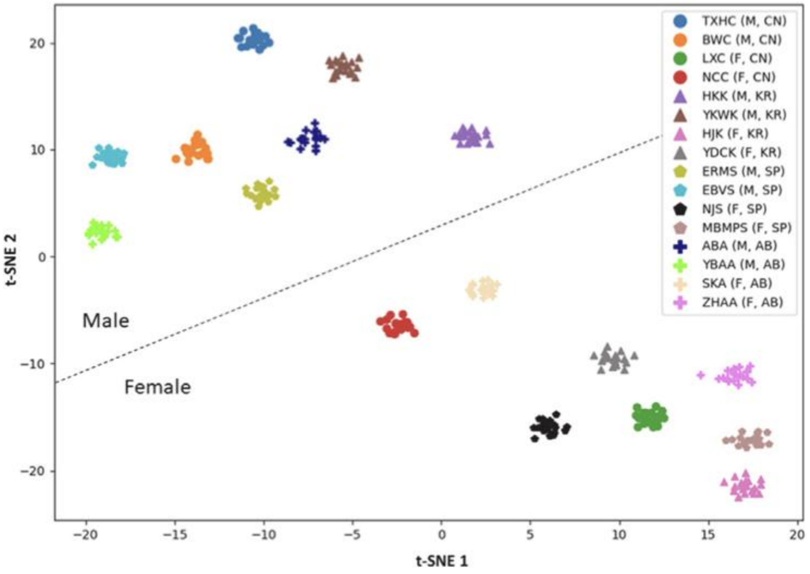


**Fig. 6.** t-SNE visualization of the speaker embeddings from 16 speakers with 4 accents. Colors and shapes represent speaker and accent, respectively. Speakers in the legend are annotated with gender and accent. L1: native accent; SP: Spanish accent; CN: Mandarin accent; KR: Korean accent; AB: Arabic accent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

# Comparison to baseline

- Accentedness
- Acoustic quality
- Voice identity

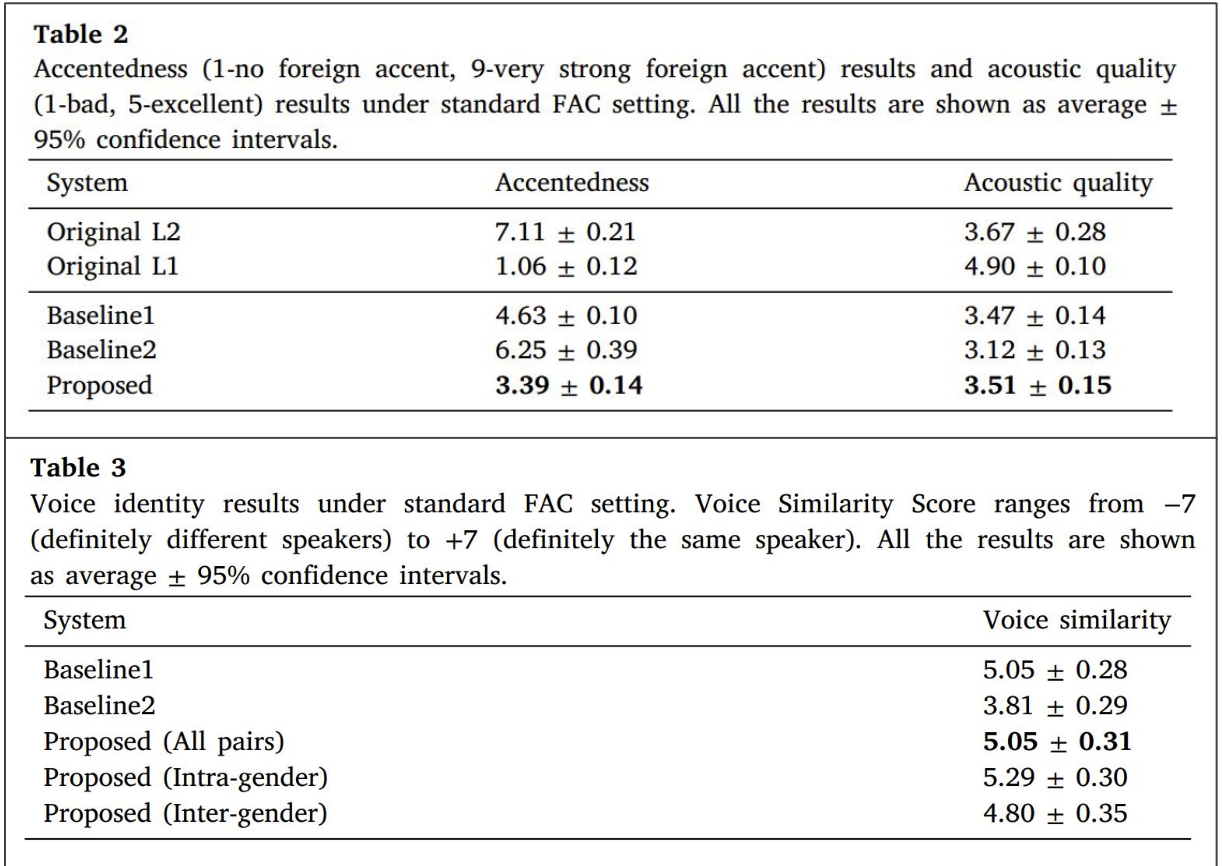


**Table 2**
Accentedness (1-no foreign accent, 9-very strong foreign accent) results and acoustic quality (1-bad, 5-excellent) results under standard FAC setting. All the results are shown as average ± 95% confidence intervals.

| System | Accentedness | Acoustic quality |
|---|---|---|
| Original L2 | 7.11 ± 0.21 | 3.67 ± 0.28 |
| Original L1 | 1.06 ± 0.12 | 4.90 ± 0.10 |
| Baseline1 | 4.63 ± 0.10 | 3.47 ± 0.14 |
| Baseline2 | 6.25 ± 0.39 | 3.12 ± 0.13 |
| Proposed | **3.39 ± 0.14** | **3.51 ± 0.15** |

**Table 3**
Voice identity results under standard FAC setting. Voice Similarity Score ranges from −7 (definitely different speakers) to +7 (definitely the same speaker). All the results are shown as average ± 95% confidence intervals.

| System | Voice similarity |
|---|---|
| Baseline1 | 5.05 ± 0.28 |
| Baseline2 | 3.81 ± 0.29 |
| Proposed (All pairs) | **5.05 ± 0.31** |
| Proposed (Intra-gender) | 5.29 ± 0.30 |
| Proposed (Inter-gender) | 4.80 ± 0.35 |

# Performance on reverse FAC

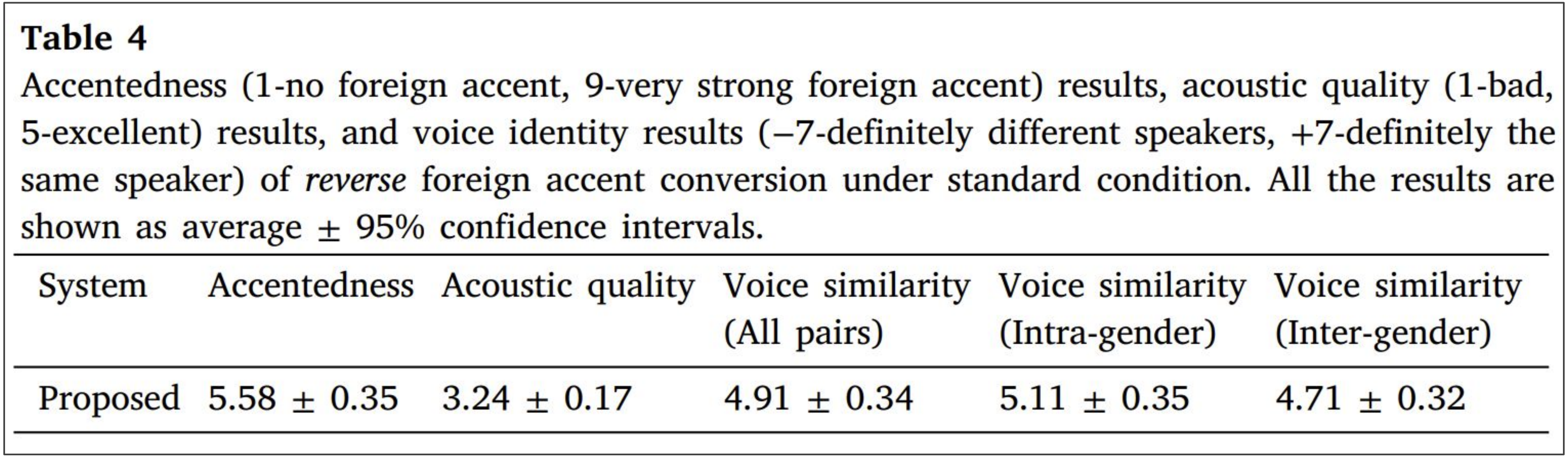


**Table 4**

Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (−7-definitely different speakers, +7-definitely the same speaker) of *reverse* foreign accent conversion under standard condition. All the results are shown as average ± 95% confidence intervals.

| System | Accentedness | Acoustic quality | Voice similarity (All pairs) | Voice similarity (Intra-gender) | Voice similarity (Inter-gender) |
|---|---|---|---|---|---|
| Proposed | 5.58 ± 0.35 | 3.24 ± 0.17 | 4.91 ± 0.34 | 5.11 ± 0.35 | 4.71 ± 0.32 |

# Comparing different conditions in zero-shot foreign accent conversion

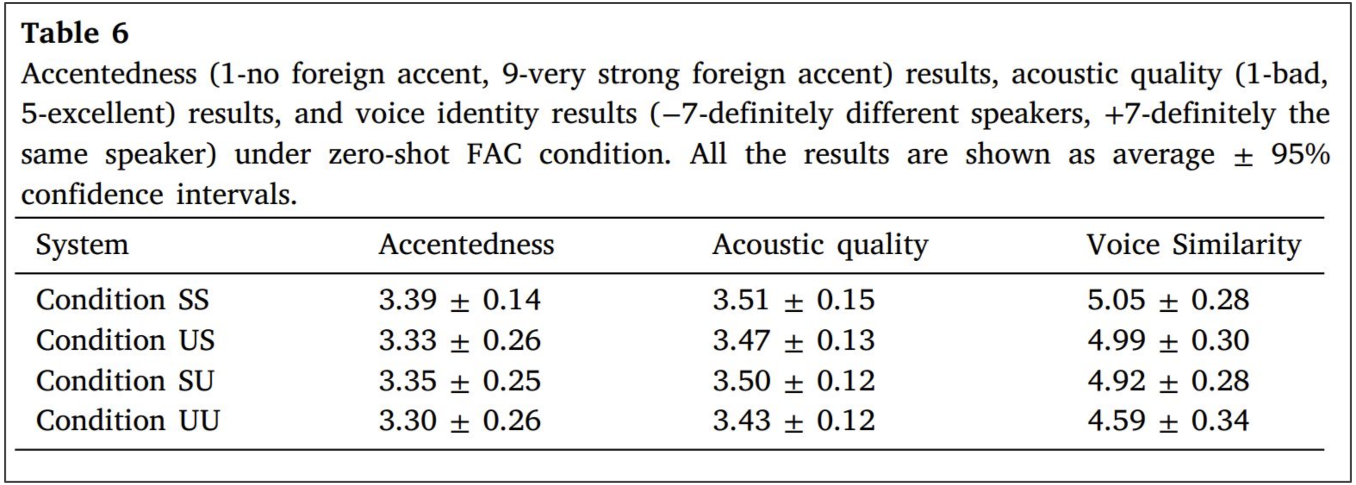


**Table 6**
Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (−7-definitely different speakers, +7-definitely the same speaker) under zero-shot FAC condition. All the results are shown as average ± 95% confidence intervals.

| System | Accentedness | Acoustic quality | Voice Similarity |
|---|---|---|---|
| Condition SS | 3.39 ± 0.14 | 3.51 ± 0.15 | 5.05 ± 0.28 |
| Condition US | 3.33 ± 0.26 | 3.47 ± 0.13 | 4.99 ± 0.30 |
| Condition SU | 3.35 ± 0.25 | 3.50 ± 0.12 | 4.92 ± 0.28 |
| Condition UU | 3.30 ± 0.26 | 3.43 ± 0.12 | 4.59 ± 0.34 |

**No Statistically significant differences** between condition SS and the three more challenging conditions

# What's the minimum amount of data needed from a target speaker?

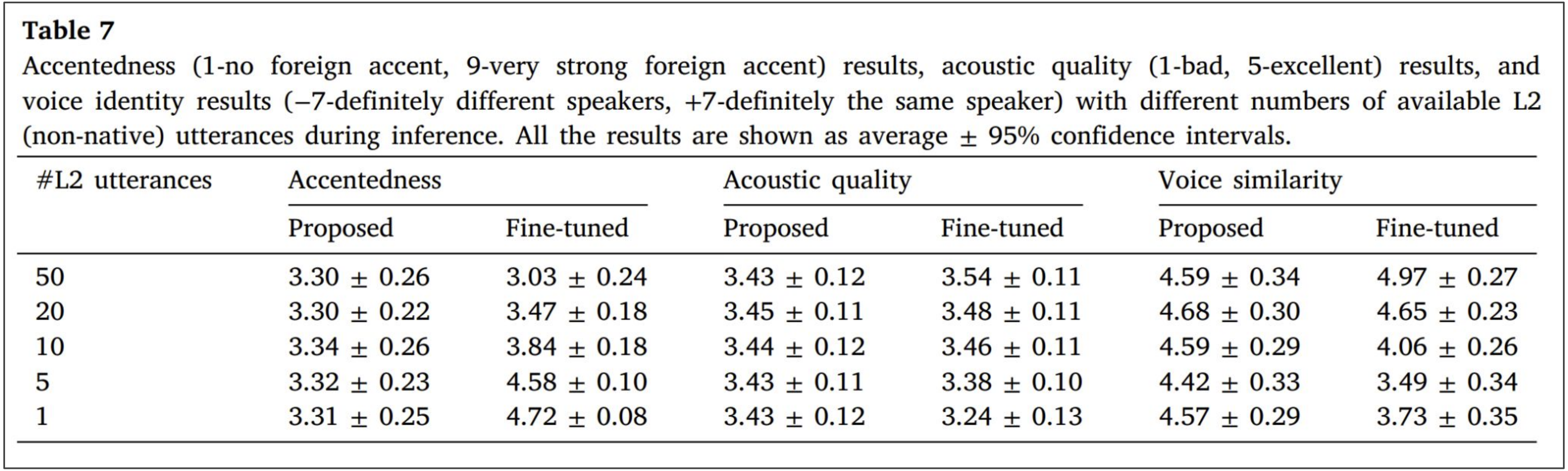


**Table 7**
Accentedness (1-no foreign accent, 9-very strong foreign accent) results, acoustic quality (1-bad, 5-excellent) results, and voice identity results (−7-definitely different speakers, +7-definitely the same speaker) with different numbers of available L2 (non-native) utterances during inference. All the results are shown as average ± 95% confidence intervals.

| #L2 utterances | Accentedness | | Acoustic quality | | Voice similarity | |
|---|---|---|---|---|---|---|
| | Proposed | Fine-tuned | Proposed | Fine-tuned | Proposed | Fine-tuned |
| 50 | 3.30 ± 0.26 | 3.03 ± 0.24 | 3.43 ± 0.12 | 3.54 ± 0.11 | 4.59 ± 0.34 | 4.97 ± 0.27 |
| 20 | 3.30 ± 0.22 | 3.47 ± 0.18 | 3.45 ± 0.11 | 3.48 ± 0.11 | 4.68 ± 0.30 | 4.65 ± 0.23 |
| 10 | 3.34 ± 0.26 | 3.84 ± 0.18 | 3.44 ± 0.12 | 3.46 ± 0.11 | 4.59 ± 0.29 | 4.06 ± 0.26 |
| 5 | 3.32 ± 0.23 | 4.58 ± 0.10 | 3.43 ± 0.11 | 3.38 ± 0.10 | 4.42 ± 0.33 | 3.49 ± 0.34 |
| 1 | 3.31 ± 0.25 | 4.72 ± 0.08 | 3.43 ± 0.12 | 3.24 ± 0.13 | 4.57 ± 0.29 | 3.73 ± 0.35 |

# Evaluation

1. Standard FAC setting
   a. Although baseline 1 and Accentron use the same backbone architecture for the seq2seq model, Accentron achieves significantly better (lower) ratings of accentedness.
2. ‘‘Reverse FAC’’ task
   a. Accentron can also preserve an L2 accent and implant it into an L1 speaker’s utterance.
3. Zero-shot FAC setting
   a. No significant differences among the four conditions in terms of accentedness, acoustic quality, or voice identity.

# Can Accentron generalize to unseen accents?

Accentron can generate FAC synthesis with high-quality, and can achieve it with limited data from new L1 and L2 speakers.

Accentron essentially reduces the resource demands and simplifying the application design and deployment.

Accentron achieves similar voice identity ratings for intra-gender pairs and inter-gender pairs.

Accentron significantly reduces the data required for each new L2 speaker from hours to seconds.

# 5. Limitations and future work

# 5. Limitations & Future works

1. From the paper:
   a. The existing system may struggle with processing when the spoken phrases are lengthy (e.g., longer than 10 s).  Possible solution: Utilizing the Gaussian mixture attention mechanism, which has demonstrated greater robustness in producing extended utterances.
   b. Incorporating an additional decoder for the purpose of phoneme recognition during the training phase.
2. From the perspective of L2 pronunciation training:
   a. Isolating and converting only certain phonetic components, e.g., vowels, consonants, prosody.
   b. Extracting linguistic content from the L2 utterance instead, thereby eliminating the need for native-accented utterances during the conversion process