Zero-Shot Voice Conversion Performance of NaturalSpeech 2 Versus Tacotron 2 and VALL-E on Low-Resource Accents

Assignee Research

June 11, 2026

Abstract

This work focuses on modelling a speaker's accent that does not have a dedicated text-to-speech (TTS) frontend, including a grapheme-to-phoneme (G2P) module. Prior work on modelling accents assumes a phonetic transcription is available for the target accent, which might not be the case for low-resource, regional accents. In our work, we propose an approach whereby we first augment the target accent data to sound like the donor voice via voice conversion, then train a multi-speaker multi-accent TTS model on the combination of recordings and synthetic data, to generate the donor's voice speaking

1 Introduction

This paper examines: Modelling low-resource accents without accent-specific TTS frontend. Research question: How does the zero-shot voice conversion performance of NaturalSpeech 2 compare to other TTS models like Tacotron 2 or VALL-E on low-resource accents when evaluated using WER and speaker similarity metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

7 papers retrieved. 20 claims extracted; 16 independently verified. Quality review score: 7.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The study models the en-IE accent using an en-GB donor speaker.	✓	0.17
The evaluation used a MUSHRA test comparing the proposed method to state-of-the-art baselines.	✓	0.22
The evaluation consisted of 100 unique testcases not seen during model training.	✓	0.18
24 native Irish speakers evaluated the testcases.	×	0.15
Systems were rated on a scale between 0 and 100 for naturalness and accent similarity.	✓	0.17
Both the reference and the hidden upper anchor in the MUSHRA test were recordings of en-IE speakers.	✓	0.17
A paired t-test with Holm-Bonferroni correction was performed to ensure statistical significance with p $\leq$ 0.05.	×	0.15
The en-IE accent differs from the en-GB accent primarily in rhoticity.	×	0.15
In en-IE, the /r/ is pronounced in postvocalic contexts when not followed by another vowel.	✓	0.21
The en-GB accent is non-rhotic.	✓	0.19
The model trained with en-GB G2P was able to reproduce rhoticity in synthesized en-IE samples.	✓	0.25
Lowering of the third formant (F3) is acoustically correlated to the phoneme /r/.	✓	0.21
An additional model was trained using phonemes extracted with the en-US G2P, which is a rhotic accent.	✓	0.23
Phoneme sequences were aligned using Dynamic Time Warping (DTW) with a cost function based on phoneme similarity.	✓	0.15
The Kaldi external aligner was used to find each phoneme position in the audio file.	×	0.12
LPC analysis was used to extract the F3 for contexts where rhoticity contrast occurs.	✓	0.20
The proposed strategy achieves state-of-the-art results compared to other generative models.	✓	0.25
Low resource accents can be modelled with relatively little data without developing an accent-specific TTS frontend.	✓	0.36
Using phoneme input avoids pronunciation errors introduced by the model mislearning how to pronounce certain words.	✓	0.24
Using phoneme input allows for phonetic control of the synthesized speech.	✓	0.17

References

- http://arxiv.org/abs/2002.03562v2
- http://arxiv.org/abs/2304.09116v3
- http://arxiv.org/abs/2301.04606v1