Neural Speech Synthesis for Austrian Dialects with Standard German Grapheme-to-Phoneme Conversion and Dialect Embeddings

Lorenz Gutscher¹,², Michael Pucher¹,², Víctor García³

¹Signal Processing and Speech Communication Laboratory (SPSC), Graz University of Technology, Graz, Austria
²Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
³HiTZ Center - Aholab, University of the Basque Country UPV/EHU, Bilbao, Spain
lorenz.gutscher@ofai.at, michael.pucher@ofai.at, victor.garcia@ehu.eus

Abstract
For languages where extensive audio data and text transcriptions are available, text-to-speech (TTS) systems have showed the ability to generate speech that closely resembles natural human speech. However, the development of TTS systems for dialects and language varieties poses challenges such as limited data availability and strong regional variations. This paper presents a TTS system tailored for under-resourced language varieties spoken in Austrian regions. The system is built upon the FastSpeech 2 architecture and includes modifications to incorporate dialect embeddings for training and inference. It is demonstrated that employing dialect embeddings and a standard German grapheme-to-phoneme conversion is effective in modeling language varieties and provides means to shift a person’s spoken variety from one to another. This allows for the generation of regional standards for dialect speakers or the generation of dialect speech with the voice of a standard speaker. The findings unveil new possibilities and applications in other multilingual contexts where shared characteristics within the language or dialect embedding space can be leveraged.

Index Terms: TTS, FastSpeech, Language embedding, Dialect modeling, Under-resourced languages

1. Introduction
Text-To-Speech (TTS) systems have undergone a notable transition towards deep learning methodologies, wherein deep learning, and especially end-to-end models, have gained significant prominence, surpassing the traditional usage of Hidden Markov Models (HMMs) [1]. This shift has been motivated by the notable improvements demonstrated by deep learning models in TTS synthesis due to the availability of large training datasets and computational resources. Prominent examples for such implementations are Tacotron [2], Tacotron 2 [3], FastSpeech [4], FastSpeech 2 [5], and VALL-E [6]. To further enhance the capabilities of TTS systems, there have been efforts to incorporate intonation and emotion controllability through speaking style modification techniques, such as Global Style Tokens (GST) [7]. These techniques operate at the level of individual utterances, enabling finer control over various aspects of speech expression.

While the transfer of accents between speakers has been explored using parallel corpora in the context of English accents [8], the absence of such source-target corpora poses a challenge. When parallel corpora are unavailable, accent transfer for TTS systems can be accomplished using an encoder-decoder setup. This involves pre-training the system with non-accented speech and subsequently fine-tuning it with accentted speech. [9] proposes to train a speech encoder that maps phoneme sequences to the target speech by pre-training a TTS system with target accentted speech and updating the encoder to minimize the loss between speech embeddings and text embeddings. Visualization of the vowel space during learning and converting General American English (24 hours of speech for a single speaker) to New Zealand English (3 hours of speech for a single speaker) is presented in [10]. [11] presents a scenario for generating accented speech with 9.60 hours of recorded speech for pre-training and less than 20 minutes for the target accent.

While accents primarily involve changes in phoneme pronunciation and prosody [12], dialects encompass a wide range of linguistic variations, including phonetic, lexical, and grammatical differences. [13, 14, 15] showcase previous approaches to modeling TTS systems for Standard Austrian German (SAG) and Austrian dialects. Additionally, audiovisual speech synthesis using HMMs is described in [16]. It is important to note that the SAG refers to the standard German spoken in Austria, which differs from Standard German German (SGG), the standard German spoken in Germany [17, 18]. In the aforementioned approaches for the synthesis of the Austrian language, only the acoustic model’s performance is investigated. Either a separate step is involved to develop a Grapheme-To-Phoneme (G2P) conversion system, or full-context dialect phoneme labels are used from phonetic transcriptions. Furthermore, a near-standard orthography is employed for dialects that is readable by non-experts [13, 14, 15].

In this study, four different varieties are considered: three dialects (Viennese dialect (V), Bad Goisern (GOI), Inneregarten (IVG)), and one standard variety (Standard Austrian German (SAG), also referred to as AT in this study). VD, the Middle Bavarian GOI, and the South Bavarian IVG dialect are examples of dialects with large deviations from the standard [19, 15]. The present study demonstrates the ability of a TTS synthesis model to acquire phonetic substitutions by integrating additional dialect embeddings and utilizing an SGG G2P system for text input. The term “dialect embedding” in this study refers to a high-dimensional vector representation used to capture the linguistic characteristics and variations specific to a particular dialect. Alternatively, an internally developed SAG G2P system could be used as a reference, but using an openly accessible SGG G2P module makes the results more applicable for other varieties (e.g., standard Swiss German). The control of the target speaker’s voice and dialect is achieved through two primary components: (I) a speaker embedding (also referred to as utterance embedding in the FastSpeech 2 implementation), and (II) a dialect embedding. Speaker and dialect embeddings are extracted and trained on a per-file basis. During inference,
a reference file is used for the speaker embedding, while a per-
variety averaged dialect embedding is utilized. Training and inference are conducted using two distinct approaches: (1) trans-
scribed phoneme labels and (2) text-level processing with a gen-
eralized G2P conversion and standard German pronunciation rules. Using a general G2P conversion has the advantage that pronunciation differences can be directly learned by the system.

The main contributions of this work include:

• Development of a publicly available Austrian German TTS system.
• Dialect shifts for combinations of speakers and dialect em-
bbeddings.
• Evaluating the perceived quality of synthesized samples for Austrian varieties.
• Evaluating the perceived effects of dialect shifts.

The paper is structured as follows: Section 2 presents a de-
tailed description of the used tools and implemented adapta-
tions. Section 3 describes the dataset, the setup for experiments, and presents results from subjective and objective evaluation metrics. Section 4 concludes the findings and contributions and outlines future research.

2. Methods

FastSpeech 2 [5] is a state-of-the-art neural text-to-speech syn-
thesis system that employs a duration predictor and a non-
autoregressive vocoder. The architecture has proven to be highly effective for TTS, particularly in scenarios with limited audio resources [20]. The implementation of FastSpeech 2 proposed in [21] is chosen as the baseline (BL) model for this work. The toolkit is adapted so that it converts labels from the Speech Assessment Methods Phonetic Alphabet (SAMPA) to the Interna-
tional Phonetic Alphabet (IPA) through a lookup table and uses a pre-trained language identification system\(^2\) to extract dia-
lect embeddings that are used for training and inference. It is hypothesized, that large-language models like wav2vec [22] and wav2vec 2.0 [23], which uses 128 languages and nearly half a million hours of speech, are sufficient to cluster dialects after being fine-tuned on the task of language identification. In a multilingual context, Austrian, Dutch, English, French, Ger-
man, Italian, Spanish, Polish, and Portuguese embeddings (100 audio samples per language) are extracted from the Common-
Language corpus\(^3\). Mapping the embedding space of these Eu-
ropean languages onto a three-dimensional space using Uni-
form Manifold Approximation and Projection (UMAP) offers insights into the acoustic proximity relationships among these languages. Notably, Austrian (located at the top) is found to have its closest neighbors in German (positioned just below the top) and Dutch (found to the left), as illustrated in Figure 1. This arrangement underscores the proximity of closely related lan-
guages, such as Austrian and German or Portuguese and Span-
ish, and provides means for using these embeddings to jointly train a TTS system, as shown in Section 3.3. In Figure 2, the embeddings of four Austrian varieties are visualized in a two-
dimensional space using 100 randomly selected utterances for each variety. It is demonstrated that the language embedding acts as a dialect embedding within a language, as GOI and IVG exhibit distinct separations. Notably, AT and VD exhibit overlapping regions, while VD appears to be centrally positioned in

\(^2\)https://huggingface.co/TalTechNLP/voxlingua107-xls-r-300m-
wav2vec
\(^3\)https://doi.org/10.5281/zenodo.5036977

the embedding space. The observed overlap between AT and VD can be attributed to the fact that certain VD samples, partic-
ularly those consisting of very short utterances, may not possess enough distinctive features to be reliably distinguished from AT, and vice versa.

While a multi-lingual pre-trained duration aligner is used for initialization, both the acoustic model for mel-spectrogram generation (FastSpeech 2) and the vocoder (HiFi-GAN [24]) are trained only on the data described in Section 3.1 without the use of a pre-trained model. The acoustic models are trained for 500 thousand steps and HiFi-GAN is trained for 2.5 million steps as suggested in the framework. Standard settings are applied to the FastSpeech 2 implementation\(^4\), with the only modification being a decreased batch size of eight for the acoustic model. The main change in architecture is the entry point of the language embedding and the way that the embedding is concatenated.

3. Experiments and results

The BL system serves as a comparison for later experiments and consists of the original FastSpeech 2 framework trained on phoneme-level label files. The second system uses per-file extracted language embeddings – treating them as dialect embeddings – and concatenates the embeddings after the utterance embedding as described in Section 2. The third system uses the before mentioned adaptation (ADP), but uses text input and a German standard G2P conversion (espeak-ng) instead of phoneme labels. Examples can be found at https://sociolectix.org/ttssigul23/.

3.1. Dataset

The dataset employed in this study consists of four Austrian varieties: SAG, VD (Vienna “Wien”), GOI (Upper Austria (“Oberösterreich”)), and IVG (East Tyrol (“Osttirol”)). The data is taken from three different corpora:

1. The Goisern and Innervillgraten Dialect Speech (GIDS) corpus is a collection of audiovisual speech recordings for research purposes. It consists of a total of 7068 sentences spoken by eight speakers (4f, 4m) from two Austrian villages, Bad Goisern and Innervillgraten [16].
2. The corpus developed within the research project “Viennese Sociolect and Dialect Synthesis” (VSDS), where three synthetic voices are built [25].
3. The Wiener Corpus of Austrian Varieties for Speech Synthesis (WASS) with read speech from a total of 19 speakers of standard Austrian German (6f, 13m) [26, 27]. The reading material contains, among others, sentences from the Berlin-Marburg corpus and the Kiel corpus, resulting in a total of 8293 utterances.

After deleting erroneous recordings and splitting the dataset into training and test files, the model is trained using 2509 utterances for GOI, 2377 utterances for IVG, 5641 utterances for VD, and 9029 utterances for SAG as described in Table 1.

3.2. Mean opinion score evaluation

The subjective evaluation is done through an online listening experiment using [28] and is administered to a nearly random sample of individuals residing in Austria whose native language is German. A total of 21 participants took part in the evaluation. While participants were familiar with the German language spoken in Austria, their familiarity with specific dialects varied. The Mean Opinion Score (MOS) of the naturalness of speech samples is evaluated on a scale from 1 to 5 (1=“Sehr schlecht (bad)”, 2=“Schlecht (poor)”, 3=“Durchschnittlich (fair)”, 4=“Gut (good)”, 5=“Ausgezeichnet (excellent)”). The evaluation involves three types of stimuli: ground truth (GT) – original recordings in 48 kHz, BL – standard implementation with phoneme labels, and ADP – adapted method with phoneme labels. The test consists of 150 speech samples: 36 for GT and 57 each for BL and ADP methods. The primary objective of the test is to evaluate whether the changed architecture either preserves or diminishes the quality of the speech samples. The evaluation results for these three systems are presented in Table 2. To interpret the results of the listening experiment, a Wilcoxon signed-rank test is performed on the rating scores due to their non-normal distribution. The GT stimuli receive a rating of 3.88, indicating that participants perceive the best achievable results to be close to the “good” range. The deviation from a score of five can be attributed to the recording conditions and participants’ challenges in evaluating the naturalness of unfamiliar dialects. BL and ADP are both rated close to “fair” (BL: 2.86, ADP: 2.85), indicating a lower quality compared to GT. However, there is no statistically significant difference (p-value = 0.67) observed between BL and ADP, suggesting that the proposed method does not result in a significant decrease compared to BL.

3.3. Standard-dialect ratings

In this section, it is evaluated whether the utilization of dialect embeddings and near-standard text input effectively induces a speaker’s shift from one language variety to another. On each page of the experiment, the same participants as in Section 3.2 are provided with a reference sample and four speech samples.

<table>
<thead>
<tr>
<th>Location</th>
<th>Gender</th>
<th>Minutes</th>
<th>Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad Goisern</td>
<td>2f, 2m</td>
<td>112.6</td>
<td>2509</td>
</tr>
<tr>
<td>Innervillgraten</td>
<td>2f, 2m</td>
<td>107.9</td>
<td>2377</td>
</tr>
<tr>
<td>Viennese dialect</td>
<td>2f, 5m</td>
<td>269.2</td>
<td>5641</td>
</tr>
<tr>
<td>Standard Austrian</td>
<td>6f, 13m</td>
<td>432.5</td>
<td>9029</td>
</tr>
</tbody>
</table>

Table 1: Training data statistics.

<table>
<thead>
<tr>
<th>Method</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>3.88 ± 0.92</td>
</tr>
<tr>
<td>Baseline (phoneme labels)</td>
<td>2.86 ± 1.04</td>
</tr>
<tr>
<td>Adaptation (phoneme labels)</td>
<td>2.85 ± 1.03</td>
</tr>
</tbody>
</table>

Table 2: MOS with 95% confidence intervals.
In a Mushra-like manner, each presented page includes a designated reference sample of the speaker’s main variety (original recording) as well as a hidden reference of the targeted variety within the stimuli (original recording of the target variety, spoken by a different individual). The four samples are then rated on a scale from 1 to 5 (1=“Dialekt (dialect)”; 2=“eher Dialekt (rather dialect)”; 3=“mittel (intermediate)”; 4=“eher Hochdeutsch (rather standard)”; 5=“Hochdeutsch (standard)”)

This experimental setup is selected due to the limited number of participants and to create a manageable task that benefits from a clearly defined reference and anchor point on the rating scale. In the example of an AT speaker and a target variety of GOI, there are five stimuli present on each page, of which four are to be rated: “reference” (original recording of a GOI speaker, unrated), “stimulus 1” (AT speaker synthesized with AT embedding, rated), “stimulus 2” (AT speaker synthesized with GOI embedding, rated), “stimulus 3 – hidden reference” (original recording of GOI speaker, rated), and “stimulus 4 – lower anchor” (original recording of AT speaker, rated).

The text input is extracted from the test set of the target variety (e.g., GOI: “Wart ein wenig, ich will dir was sagen.”). Two pages (utterances) are presented for each of the dialect shifts of one standard speaker (2*1*[AT-VD, AT-GOI, AT-IVG]), one dialect speaker (2*1*[VD-AT, VD-GOI, VD-IVG]), four GOI speakers (2*4*[GOI-AT]), and four IVG speakers (2*4*[IVG-AT])

Two pages with different samples are presented to each participant with a total of 28 pages. Figure 4 shows the mean box plots of all utterances (for simplicity, AT embeddings are averaged in Figure 4a, as well as VD embeddings in Figure 4b).

Significance testing is conducted for paired comparisons, yielding the following results:

• AT: In the case of one AT speaker, a significant difference in rating exists when comparing stimuli using {AT and VD} embeddings, {AT and GOI} embeddings, and {AT and IVG} embeddings.

• VD: In the case of one VD speaker, {AT and VD} embeddings show a significant difference, but there is neither a statistically significant difference between {VD and GOI} embeddings, nor between {VD and IVG} embeddings. In the case of this particular speaker, the distinction between AT and VD is less pronounced compared to the other speakers. This observation aligns with the overlapping regions of those two varieties, as illustrated in Figure 2.

• GOI: In the case of four GOI speakers, there is a statistically significant difference between {GOI and AT} embeddings. This finding suggests that the inclusion of a limited number of utterances spoken in AT within the data for GOI (and IVG) speakers positively impacts the model’s ability to capture dialect shifts.

• IVG: In the case of four IVG speakers, there is a statistically significant difference between {IVG and AT} embeddings.

However, it is important to note that this test is intended to showcase the feasibility of shifting a standard speaker to a dialect. With this feasibility confirmed, future tests need to be specifically designed to assess whether the shifted speech is perceived as the intended target dialect and not merely any dialect.

To validate the preservation of the speaker’s voice attributes after the shift in variety, a speaker verification system is employed to measure the cosine similarity [29] between original and shifted samples. While a cosine similarity score of 1 indicates perfect similarity between two speakers, a score close to 0 signifies dissimilarity, i.e., different voices. Using two original recordings (R1, R2) as reference samples for each speaker and one synthesized shifted sample (S) from the same speaker, the cosine similarity is calculated between R1-R2 (reference value), R1-S, and R2-S. Calculations are done for each shifted example that is presented in the listening experiment. The average similarity score for R1-R2 samples over all speakers is 0.81, while the average score of R1-S and R2-S is 0.79, indicating that the shifted samples originate from the same speaker as the references.

### 4. Conclusions

This paper presents an effective method to incorporate dialect embeddings for training a FastSpeech 2 text-to-speech synthesis model. It was shown that dialects can be effectively modeled using near-standard orthography and that the spoken language variety of a speaker can be shifted towards standard or dialect without changing the speaker similarity. This enables, e.g., the generation of region-specific standard varieties for dialect speakers and facilitates smooth interpolations between different dialect varieties. To further validate the authenticity of the shifted dialect, future work is going to involve a phonetic analysis of synthesized speech samples. This analysis will involve a comparison between samples of the shifted dialect and samples of a GT speaker who is native to the dialect.

![Figure 4: Subjective standard-dialect ratings using standard and dialect embeddings.](image-url)
5. References


