Unify and Conquer: How Phonetic Feature Representation Affects Polyglot Text-To-Speech (TTS)

Ariadna Sanchez, Alessio Falai, Ziyao Zhang, Orazio Angelini, Kayoko Yanagisawa

Alexa AI, Amazon
{carladn,falai,zhaziyao,aorazio,yakayoko}@amazon.com

Abstract
An essential design decision for multilingual Neural Text-to-Speech (NTTS) systems is how to represent input linguistic features within the model. Looking at the wide variety of approaches in the literature, two main paradigms emerge, unified and separate representations. The former uses a shared set of phonetic tokens across languages, whereas the latter uses unique phonetic tokens for each language. In this paper, we conduct a comprehensive study comparing multilingual NTTS systems models trained with both representations. Our results reveal that the unified approach consistently achieves better cross-lingual synthesis with respect to both naturalness and accent. Separate representations tend to have an order of magnitude more tokens than unified ones, which may affect model capacity. For this reason, we carry out an ablation study to understand the interaction of the representation type with the size of the token embedding. We find that the difference between the two paradigms only emerges above a certain threshold embedding size. This study provides strong evidence that unified representations should be the preferred paradigm when building multilingual NTTS systems.

Index Terms: neural text-to-speech (NTTS), multilingual speech synthesis, polyglot TTS, linguistic features

1. Introduction
Text-To-Speech (TTS) has been attempting to move from monolingual to multilingual applications for a long time [1]. Advancements in Neural TTS (NTTS) now allow us to create high-quality multilingual and multi-speaker speech synthesis models [2, 3]. Multilingual TTS models differ from each other significantly in terms of the level of flexibility of speaker/language combinations that they are able to deliver. Some models rely on bi-/multi-lingual recordings from the same speaker to generate multilingual synthesis [4, 5], i.e. speaker and language are coupled. Other multilingual models are capable of generating speech in languages for which there are no available recordings from the speaker [6]. In this paper, we give the name Polyglot TTS to the multi-speaker multilingual TTS system that enables any speaker in the training corpus to speak any language present in it.

From a linguistic and design perspective, modelling input features for Polyglot TTS systems presents an interesting research question. Previous Polyglot TTS work has used various input representations. We divide them into two main groups: unified representations, which are normally based on the International Phonetic Alphabet (IPA) [7], and separate representations, which entail having separate input tokens for each language. For unified representations, there is previous work based on the IPA features [8, 9]. Other works used phonetic features extracted from dictionaries like CMUDict [10], or using a hand-crafted phone-set [11], which is not a scalable approach. Some previous work also defines phones with phonological features instead of using a single token per phoneme, thus allowing for zero-shot multilingual synthesis [12, 13]. There are other unified representation approaches which, instead, have used text-based features such as characters [14] or bytes [15, 16].

For separate representations, [2] studies the use of graphemes, byte-encoding and phonemes by concatenating symbols from all the language vocabularies, with common tokens shared between them. Another work [17] investigates learning phonetic mappings between two languages. In [18], authors use the same phonetic representation but separate phoneme encoders for each language.

Conceptually, there are arguments in favour of both paradigms. For example, having separate tokens per language and accent may help the model avoid accent leakage, i.e. the synthesis having a slight accent from the original language of the target speaker, because all phonetic tokens are unique. On the other hand, we can also hypothesise that a language-agnostic solution can improve the accent and naturalness by removing redundant tokens and thus allowing the model to generalise over similarities between languages and accents.

The main contribution of this paper lies in the experimentation and evaluation aimed at understanding how unified representations and separate representations of input linguistic features affect the quality of Polyglot synthesis in terms of naturalness and accent of the voice. To the best of our knowledge, this is the first work conducting a systematic study and evaluation on this subject.

2. Methodology

2.1. Polyglot Text-to-Speech

2.1.1. Architecture
Our base architecture consists of a neural acoustic model and vocoder pair. The acoustic model, tasked to convert phonetic sequences into mel-spectrograms, follows [19], which is a sequence-to-sequence attention-based model [20]. We use a universal vocoder based on [21], which takes as input mel-spectrograms and outputs raw waveforms.

The acoustic model is composed of 3 major components: an encoder, a decoder and an attention mechanism. Phonemes are given as input to the encoder, which processes them through a stack of 3 convolutional blocks, followed by a single bi-LSTM unit. The decoder comprises a pre-net and a stack of 2 LSTM layers, to then branch out into mel-spectrogram and stop-token heads, in order to predict the mel-bands for the current time-frames and the probability of cutting off auto-regressive generation, respectively. A final post-net acts as a residual signal to adjust mel-spectrogram predictions. In order to learn how many frames correspond to each input phoneme, we rely on a location-sensitive cell based on [22], with the same settings as
Multi-speaker, cross-lingual synthesis is enabled with lookup tables pointing at 256-dimensional trainable embeddings to condition the encoder on speaker, language and language variant (LV) (refer to Table 1). These embeddings are fed to the encoder and the attention mechanism, as shown in Figure 1. To condition the model on style\textsuperscript{1}, we rely on a reference encoder [24] composed of a Variational Auto-Encoder (VAE) followed by 16 

**Figure 1: Architecture used for the acoustic model, explained in Section 2.1.1. For some models, language conditioning is not used, as explained in Section 3.**

reported in [23].

At inference time, we are able to synthesise speech with the desired speaker, language, LV and style, by conditioning the model on the appropriate embeddings. For speaker, language and LV, we simply select representations from each lookup table, while for style we use a pre-calculated centroid of the available ground truth data for the target speaker in the target style, in place of the VAE-generated embedding.

All models are trained with the Adam optimiser and a combination of $L_1$, Kullback–Leibler Divergence (KLD) and Cross Entropy (CE) loss functions for spectrogram, VAE and stop token, respectively, similarly to [26]. We also make use of an exponentially decaying learning rate, that starts at $10^{-3}$ and is reduced by 2% every 10k steps, and follow the standard teacher forcing [27] approach to predict 5 mel-spectrogram frames at each iteration. Hyper-parameters (such as the learning rate schedule) are selected following standard practices in the literature [28, 23]. Their particular values marginally affect the final results, but not the conclusions.

### 2.1.2. Data composition

Our dataset consists of 130 speakers from 12 LVS across 4 languages, which are summarised in Table 1. LVS belong to either the Romance or Germanic language family. The dataset contains between 3 and 6 hours of recordings per speaker. The target speaker, i.e. the speaker for which we want to enable cross-variant synthesis, has instead more than 30 h of recordings. We do not impose a uniform balance between the amount of data in each language and LV, as it is beyond the scope of this paper. Overall, the dataset is composed of 580 h of recordings.

<table>
<thead>
<tr>
<th>Language</th>
<th>Language variant (LV)</th>
<th>Language variant code</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Australia</td>
<td>en-AU</td>
</tr>
<tr>
<td></td>
<td>India</td>
<td>en-IN</td>
</tr>
<tr>
<td></td>
<td>Wales</td>
<td>en-WLS</td>
</tr>
<tr>
<td></td>
<td>Britain</td>
<td>en-GB</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>en-US</td>
</tr>
<tr>
<td>French</td>
<td>Canada</td>
<td>fr-CA</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>fr-FR</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>es-US</td>
</tr>
<tr>
<td>Spanish</td>
<td>Spain</td>
<td>es-ES</td>
</tr>
<tr>
<td></td>
<td>Mexico</td>
<td>es-MX</td>
</tr>
<tr>
<td>German</td>
<td>Germany</td>
<td>de-DE</td>
</tr>
</tbody>
</table>

### 2.2. Phonetic representations

This paper compares two different input representations, which we call (language-)separate and (language-)unified, and are defined as follows.

- **Separate:** In this paradigm, there is a separate phone-set for each LV. In each phone-set, each phoneme has its own unique token. In other words, even phones which share the same symbol in different LVS have separate tokens in the input representation (e.g. /b/ in American and British English or /a/ in Canadian French and German).

- **Unified:** We use a unified representation of phones for all LVS in the training data. In this paradigm, the phones are denoted using X-SAMPA [29], a machine-readable notation for the IPA. The same set of tokens is used across all LVS, e.g. /b/ in American English and British English correspond to the same token.

In *unified* features, both service tokens (such as word boundary and utterance start tokens) and punctuation tokens are shared across languages, just like the phoneme tokens. On the other hand, *separate* representation models are trained with shared service tokens, but punctuation tokens are separate for each language, like the phoneme tokens. The reason for this is explained in Section 3.1.

### 3. Experiments

All subjective evaluations conducted are controlled in terms of training data compositions and the number of training epochs. A slightly different architecture is used for the 2 input paradigms. For *separate* features, we condition the model on speaker, language and LV; while for *unified* features, we only use speaker and LV conditioning. Subjective evaluations revealed that, when using *separate* features, conditioning on speaker, language and LV is significantly better than conditioning only on LV. Contrarily, for *unified* features, conditioning only on speaker and LV leads to significantly better accent, while also adding language conditioning brings accent degradation caused by the target speaker’s own accent. We hypothesise that, for *unified* features, when the model sees both language and LV conditioning, it diminishes their importance. This,
in turn, makes the model more reliant on the speaker embedding, which contains language information of the target speaker, hence impacting the accent negatively.

3.1. Comparison of phonetic representations

We compare separate and unified phonetic representations by training two models, one for each kind of input phonetic features. For both models, the target LVs we evaluate are de-DE, fr-CA, en-GB and es-US.

We compare each model for naturalness and accent separately, using a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [30]. For each of the target LVs, we synthesise 100 samples from both models for evaluation. A total of 60 listeners participate in each evaluation, who listen to 50 randomly selected utterances from the pool of 100 samples. For preserving the integrity of our evaluations, we use upper and/or lower anchors as control samples. We filter out any listeners that consistently give undue high/low scores to the lower/upper anchors, respectively. No more than 6 listeners were excluded from each evaluation.

The systems included in naturalness evaluation are (1) separate model, (2) unified model, and (3) the original en-US target speaker recordings, which serve as the upper anchor. We ask the listeners to rate how natural each sample sounds on a scale of 0 (“completely unnatural”) to 100 (“completely natural”), ignoring any differences in language or accent. Figure 2 shows the results for each LV, where synthesis with unified features scores higher than with separate features, with p-value $\ll 0.001$ for all comparisons.

The systems included in accent evaluation are (1) separate model, (2) unified model, (3) the recordings from a native speaker of the target language as upper anchor, and (4) the synthesis from a model that has been trained with monolingual data of the target speaker (an American English speaker) as lower anchor. For the lower anchor, we map each phone of the target language with a similar phone that exists in the monolingual model, purposely producing American English-accented speech. We ask each listener to rate how native does each sample sound from 0 (“completely foreign”) to 100 (“completely native”). Figure 3 shows the MUSHRA results for each LV. Synthesis with unified features scores higher than with separate features in each case, with p-value $\ll 0.001$ for all comparisons, although the gaps between the mean values are smaller.

We have typically observed that, when a model is trained with multilingual corpora, there is quality degradation in the target speaker’s synthesis in their native LV, which may be due to training data composition [31]. Therefore, we also evaluate the models in terms of naturalness, accent and speaker similarity for the native LV of the target speaker, i.e., en-US. Naturalness and accent are evaluated via MUSHRA tests. Following [32], we use the pre-trained speaker verification model Resemblyzer\(^2\) to measure speaker similarity. Figure 4 shows that the difference in both accent and naturalness between our proposed models and the monolingual model is not statistically significant. This suggests that both input features result in audio samples of comparable quality to that of the monolingual model when speaking the target speaker’s native LV.

We additionally observe that the unified representation model is more stable than the separate representation model,

\[^2\]https://github.com/resemble-ai/Resemblyzer
with fewer occurrences of mumbling or skipped punctuation pauses. We hypothesise that reducing the number of phonetic tokens correspondingly reduces the number of contexts the model has to learn, while keeping the amount of data constant, i.e., the model sees more examples of each possible context and is poised to generalise better. This enables the model to learn a more robust representation of the phonetic features, leading to a more stable model. Sharing the punctuation tokens across LVs also helps to increase the amount of seen data for each punctuation, leading to better pausing. In the case of separate features, merging the punctuation tokens across languages led to instabilities. We suspect that this is because for a given punctuation token (e.g. comma), there is around one order of magnitude more possible surrounding phonetic contexts compared to the unified case, since there are more phonetic tokens. This, in effect, reduces the model’s ratio of observed versus possible contexts, leading to instabilities. Hence for the separate case we kept the punctuation tokens separate for each LV.

3.2. Ablation study for phoneme embedding size

The number of input tokens in the separate and unified paradigms differs significantly. We therefore investigate the effect of the size of the phoneme embedding vector in both cases. In this section we describe an ablation study where the size of the phoneme embedding \( s \) is gradually reduced. This would confirm whether the results in Section 3.1 still hold when the phoneme embedding size changes. It also reveals the minimum \( s \) to avoid synthesis quality degradation.

We train Polyglot models with separate and unified features using \( s = 64 \) and \( s = 128 \), while keeping all other parameters fixed. We then compare such reduced capacity models between themselves and against the ones reported in Section 3.1, for which \( s = 256 \).

Specifically, we run 2 different sets of MUSHRA tests for naturalness, accent and one comparing separate and unified features with each other at lower embedding sizes (results in Table 4); and another to identify the best \( s \) parameter for both separate and unified features (results in Table 3).

### Table 3: Mean MUSHRA scores relative to recordings for naturalness (N) and accent (A), comparing different phoneme embedding sizes within the separate and unified features, across all evaluated LVs.

<table>
<thead>
<tr>
<th></th>
<th>Separate</th>
<th>Unified</th>
<th>Separate</th>
<th>Unified</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>83.40</td>
<td>82.79</td>
<td>78.70</td>
<td>74.99</td>
</tr>
<tr>
<td>128</td>
<td>92.15</td>
<td>87.65</td>
<td>89.25</td>
<td>90.29</td>
</tr>
<tr>
<td>256</td>
<td>78.00</td>
<td>79.65</td>
<td>85.66</td>
<td>83.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Separate</th>
<th>Unified</th>
<th>Separate</th>
<th>Unified</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>82.97</td>
<td>82.94</td>
<td>84.17</td>
<td>82.83</td>
</tr>
<tr>
<td>128</td>
<td>92.44</td>
<td>89.19</td>
<td>90.91</td>
<td>92.57</td>
</tr>
<tr>
<td>256</td>
<td>88.15</td>
<td>86.04</td>
<td>91.05</td>
<td>89.47</td>
</tr>
</tbody>
</table>

Interestingly, results clearly indicate that for unified features, bigger phoneme embedding sizes are consistently preferred over smaller ones for both naturalness and accent for all LVs. Such results are statistically significant with all p-value \(< 0.005\). This seems to suggest that for unified features, more phonetic information is captured in the phoneme embedding vectors, making higher dimensional vectors beneficial. In contrast, for separate features, no particular patterns were detected, which lead us to conclude that \( s = 64 \) is sufficient in this case. We hypothesise that the difference in results is explained by two factors. One is the increased total number of phonemes in the separate features, which by itself increases the capacity of the model. The other is the reduced amount of times a phoneme is observed when switching from separate to unified features. This allows each embedding to learn less information. The direct comparison between separate and unified features (Table 4) confirms instead that the conclusions in Section 3.1 are affected by the phoneme embedding size. In particular, we find that the differences between separate and unified features are less marked at lower embedding sizes, where the clear signal from Section 3.1 disappears. Note in fact that the differences in relative mean scores observed between the two feature types are not statistically significant, for both naturalness and accent. In summary, the benefits of unified representations start to manifest themselves only above a certain threshold of phoneme embedding size, while the quality of separate representation models is not affected by a variation in phoneme embedding size.

### Table 4: Mean MUSHRA scores relative to recordings for Naturalness (N) and Accent (A), comparing separate and unified features at lower phoneme embedding sizes, across all evaluated LVs.

<table>
<thead>
<tr>
<th></th>
<th>Separate</th>
<th>Unified</th>
<th>Separate</th>
<th>Unified</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>77.03</td>
<td>77.50</td>
<td>80.44</td>
<td>76.52</td>
</tr>
<tr>
<td>128</td>
<td>90.12</td>
<td>87.88</td>
<td>90.11</td>
<td>93.20</td>
</tr>
<tr>
<td></td>
<td>81.31</td>
<td>84.48</td>
<td>80.17</td>
<td>79.54</td>
</tr>
<tr>
<td></td>
<td>92.89</td>
<td>89.92</td>
<td>100.40</td>
<td>100.13</td>
</tr>
</tbody>
</table>

4. Conclusion

In this work, we showed that a unified phonetic representation can substantially improve the accent and naturalness of a Polyglot TTS model. Despite the phonetic representation being shared across all LVs, the model did not suffer from more accent quality degradation compared to the separate representation model. Speaker similarity to the target speaker is also significantly improved compared to the model trained with separate features, bringing it closer to that of single-speaker monolingual models. We also observed that the unified representation model was more stable than the separate representation model, likely due to a reduced number of tokens which allows for better generalisation over phonetic similarities between languages and their variants. Furthermore, we showed that the difference between unified and separate features only emerges at larger embedding sizes. As an extension of this work, we could consider extending this work to languages from other language families. Moreover, we could also consider a comparison with representations based on phonological features. These features may allow the model to learn about similarities between phonemes in different languages represented by different symbols.
5. References


