Continuous features in neural TTS

Christina Tånnander[1,2], Jim O'Regan[1], Shivam Mehta[1], David House[1], Jonas Beskow[1], Jens Edlund[1]

[1]Royal Institute of Technology
Stockholm, Sweden

[2]Swedish Agency for Accessible Media
Malmö, Sweden

Abstract

This paper describes work in which we go beyond conventional discrete representations of graphemes and phonemes in TTS training and instead use continuous feature value vectors as inputs for neural text-to-speech (TTS) synthesis. The two studies described demonstrate the potential of continuous features in controlling various phonetic-phonological aspects of synthetic speech. The first study introduces a continuous language feature to modulate the degree of English-accentedness in Swedish speech synthesis. The second study employs continuous phonological features to represent American English speech sounds. Our findings indicate that continuous feature representations can enhance the flexibility of neural TTS systems, with potential applications in multilingual speech synthesis, accented speech generation, and synthesis for under-resourced languages. In other words, the work paves the way for more versatile TTS systems as well as improved opportunities for research based on analysis-through-synthesis.

1 Introduction

Input to neural TTS (text-to-speech synthesis) typically consists of discrete grapheme or phoneme representations of the text. While graphemes are common in the TTS research community, where the goal often is to try out new machine learning approaches, industrial approaches often use phoneme input to ensure better control over the wording and pronunciation (see e.g. Acapela Group, 2005; CereProc Ltd, 2023; Google, 2023; Microsoft, 2022).

Several studies have explored the consequences of using graphemes, phonemes (e.g. Fong et al., 2019; Taylor et al., 2021), or even phonological features as input to neural TTS (e.g. Staib et al., 2020; Maniati et al., 2021). In this work, we take the feature approach one step further, and use *continuous* features vectors, each feature holding a value between 0.0 and 1.0. As a result, we can assign intermediate values to achieve more fine-grained, gradual control over the feature. For example, a nasality feature, which in traditional theory is a binary feature and takes the values + or -, can here take any value between 0.0 and 1.0. In a study using the OverFlow voice presented below, this nasality feature was successfully varied over full utterances (Näslund et al., 2024).

Here, we present two previously published studies in which we successfully control features by degree in neural TTS. The first study concerns a continuous *language* feature, making it possible to synthesise Swedish with an increasing degree of English-accentedness (Tånnander et al., 2024a). During training, Swedish phonemes were assigned a language feature value of 0.0 (no accent) and English phonemes were assigned 1.0 (full English accent). The test data was then synthesised using intermediate accent values.

The second study used 11 continuous *phonological* features to represent American English. Here, some of the features also involved intermediate values in training, as well as during inference. Two features were evaluated with categorical perception tests and acoustic analyses (Tånnander et al., 2024b).

2 Method

Continuous feature values ranging from 0.0 to 1.0 were used in both studies. Note that the feature values do not represent absolute values but should be regarded as rank orders. For example, a nasality value of 0.25 does not correspond to 25 % nasality but is expected to be more nasal than 0.0 and less nasal than 0.5

2.1 Continuous language feature

A continuous language feature representing different degrees of English-accentedness in Swedish was introduced. For further details about the study, see Tånnander et al. (2024a).

TTS framework: The voice was trained using OverFlow (Mehta et al., 2023). Hifi-GAN (Kong et al., 2020) was used as vocoder.

Training: Around 12 000 Swedish and 8 000 English sentences read by the same professional Swedish, female speaker (Tånnander, 2018). The language feature was set to 0.0 for Swedish and to 1.0 for English speech sounds.

Test data: The Swedish translation of *The North wind and the sun* ('Swedish Phonology', 2023) and four constructed Swedish sentences were used as test data. Five degrees of the language feature (referred to as TEA, Targeted English-Accentedness) were evaluated, the extremes 0.0 and 1.0, which were expected to produce Swedish without any accent and with full English accent, as well as three intermediate values (Figure 1).

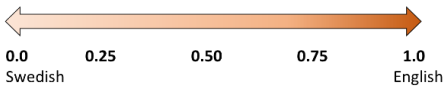

Figure 1. The five degrees of English-accentedness used in the test stimuli.

Test procedure: *Speech intelligibility* was evaluated using the output word error rates (WER) from the speech recognition module of Whisper (Radford et al., 2022).

English-accentedness was evaluated automatically and through a perception test. The automatic verification used Whisper's language classification module, where the probability of the speech being English was hypothesised to increase with higher targeted degrees of English-accentedness. In the perception test, 20 subjects listened to sentence pairs with different degrees of English-accentedness and selected the most English-accented rendition.

In addition, *durational and F0 analyses* using REAPER (*REAPER*, 2014/2023) aimed to explore differences between non-accented and English-accented phoneme durations.

2.2 Continuous phonological features

Each speech sound was represented by a vector of 11 continuous feature values. Similar to the language feature in the previous study, each feature took any value between 0.0 and 1.0, but in this case, intermediate values were also used for phonological features during training. The study is described in more detail in Tånnander, et al. (2024b).

R	C-PLACE	Example	V-HEIGHT	Example
1	glottal	/h/	most open	/æ/
2	velar	/k/		/ɛ/
3	palatal	/j/		/ɜː/
4	post-alveolar	/ʃ/		/ɪ/
5	alveolar	/t/		/ʊ/
6	dental	/θ/	most close	/iː/
7	labiodental	/f/		
8	bilabial	/p/		

Table 1: The training rank order (**R**) of **C-PLACE** and **V-HEIGHT**, mapped to place of articulation and relative vowel height.

TTS framework: The voice was trained using Matcha-TTS (Mehta et al., 2024). Hifi-GAN (Kong et al., 2020) was used as vocoder.

Training: The training data consisted of an American English speech database: RyanSpeech, with almost 10 000 American English sentences (Zandie et al., 2021). 11 phonological features were used. 5 were assigned 0.0 or 1.0 values only (e.g. **VOICING**), while 6 also included intermediate values during training (e.g. vowel tongue positions and consonant stricture. The evaluation concerns the two features discussed in more detail here: **C-PLACE** and **V-HEIGHT**. **C-PLACE** is related to the place of articulation, with glottal at the lower and bilabial at the higher end of the continuum (see Table 1). As mentioned above, the intermediate numbers should be interpreted as rank orders rather than exact positions. They do not correspond to for example physical distances in the vocal tract; (2) velar is just closer to the beginning of the continuum than (4) post-alveolar is. Similarly, **V-HEIGHT** shows the relative vertical position of the tongue; a low **V-HEIGHT** value shows a vowel that is more open than a higher value. Figure 2 uses the phoneme /ɛ/ to illustrate the association with traditional, binary phonological features.

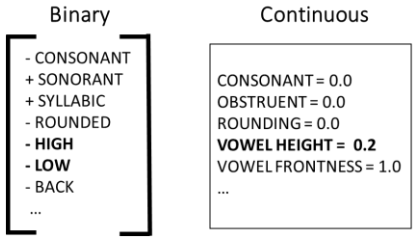

Figure 2: Excerpts of binary phonological features and continuous features of /ɛ/.

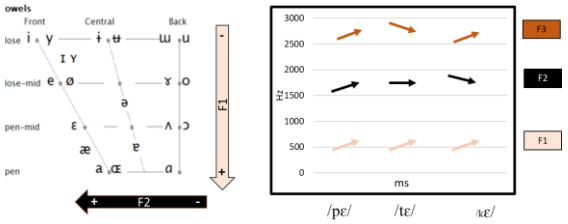

Figure 3: Left: the relation between F1 and vowel height. Right: expected F1-F3 slopes when /ɛ/ follows a bilabial, alveolar or velar stop.

Test data: Stimuli where only one focus feature was altered involved nasals and voiceless stops on the **C-PLACE** continuum and front vowels on the **V-HEIGHT** continuum. The target words were embedded in the carrier sentence *I say <target_word> again*. Listen to the test stimuli at https://www.speech.kth.se/tts-demos/interspeech24phonological/.

Test procedure: To verify that the features could be controlled by degree, categorical perception tests and acoustic analyses were performed. In the categorical perception tests, 120 subjects marked whether they perceived *wi**m***, *wi**n*** or *wi**ng*** for stimuli with a nasal at 9 different locations on the **C-PLACE** continuum; ***p**en*, ***t**en* or ***k**en* were used as corresponding voiceless stops; and *b**i**t*, *b**e**t* and *b**a**t* for the front vowels at 9 different locations on the **V-HEIGHT** continuum.
The acoustic analyses concerned F1 values for the **V-HEIGHT** altered vowels, where F1 is expected to increase with lower rank order (more open vowel). For **C-PLACE**, the F2 slope over around 50 ms transitions between target consonants and vowels was measured. The F2 transition is expected to rise when /ɛ/ follows a bilabial phoneme, to be at level after alveolars, and to fall after velars. These expectations are illustrated in Figure 3.

3 Results

3.1 Continuous language feature

The speech intelligibility Whisper test showed a WER starting at under 10% at 0.0, and then slightly increasing with higher TEA (Targeted English-Accentedness, see Figure 4). Whisper's language classifier resulted in probabilities over 99% for Swedish at 0.0-0.50, and English probabilities increasing with higher TEA, although the speech was correctly identified as Swedish also at a TEA of 1.0 (p=>0.96).

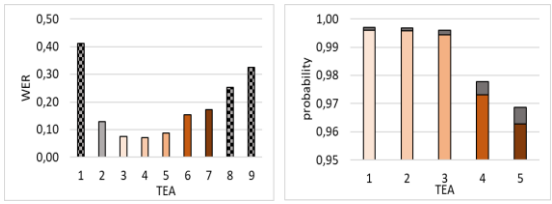

Figure 4: Left: WER for varying TEA. Right: Estimated probability of Swedish (lower) and English (upper) (truncated x axis). s0 represents 0% TEA, and s100 100% TEA.

The pairwise perception test, where one utterance always had a higher intended TEA than the other, showed that the listeners selected the utterance with the highest TEA in 89% of the cases. These results show that the synthesised speech is intelligible, and that he intended degree of English-accentedness was achieved.

Durational measurements showed a strong correlation between Swedish/English recorded sentences and the synthesised test material, with shorter vowel durations in English and English-accented speech as its main finding. Finally, f0 in English-accented synthesised speech was generally lower than in non-accented sentences.

3.2 Continuous phonological features

The categorical perception test, where subjects listened to 9 degrees of **C-PLACE** or **V-HEIGHT** and selected which of three words they perceived, showed that they chose the option closest to the targeted feature. Figure 5 illustrates the perceptual categorizations as /k/, /t/ or /p/ for the 9 tested values of **C-PLACE**. We see that the black F2 and orange F3 arrows approximately correspond to the expected slopes of the formant transitions shown in Figure 3.

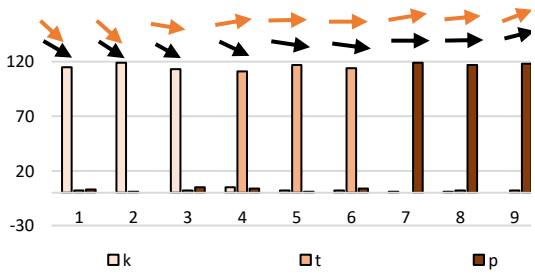

Figure 5: Count of perception of /p, t, k/ (y-axis) for 9 stimuli (x-axis) of **C-PLACE** stops before vowel. Black arrows illustrate F2 slope and orange arrows F3 slope.

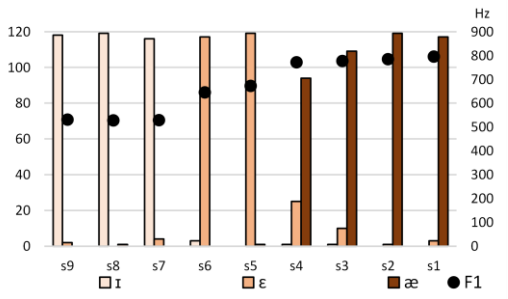

Figure 6: Counts of perception of /ɪ , ɛ, æ/ (left y-axis) for each of the 9 settings (x-axis) of **V-HEIGHT**. Training target phonemes labels are aligned with their rank order. Average F1 value is shown as black dots (right y-axis).

The **V-HEIGHT** perception test showed similar tendencies; the closest vowel option was generally chosen, although there was some confusion at rank order (4), as shown in Figure 6. The average F1 values are here shown as black dots.

4 Discussion

The experiments showed that representing phonological features as continua in the training enables more refined control over the speech generation compared to traditional discrete features. When using the same targets as were used in training for generation, we could hear no noticeable loss of quality compared to discrete features.

The continuous *language* feature investigated in the first study controls the degree of English-accentedness in synthesised Swedish speech, with applications ranging from speech science (with more thorough validation, the synthesised speech could be used as material for analyses of English-accentedness), to better renditions of for example new loan words, English names, and English embedded phonemes in general. The approach successfully produced variations ranging from no accent to a fairly full English accent, with intermediate levels effectively perceived as such, both by automatic systems and human listeners.

The continuous *phonological* features representing American English speech sounds in the second experiment gave the ability to produce smooth transitions and gradual changes in speech sounds, validated in categorical perception tests and acoustic analyses.

We conclude that continuous feature representations provide a flexible framework for capturing the subtle gradations of speech sounds. This flexibility is particularly beneficial for tasks that require nuanced control over speech output, such as generating accented speech or blending features from different languages.

These findings open promising avenues for future speech science, both fundamental and applied. One significant area is the development of multilingual and accented speech synthesis systems capable of code switching both within and between utterances. TTS systems driven by continuous features can more seamlessly switch between languages and accents, providing a more realistic and cohesive user experience. This approach also holds potential for the customisation and personalisation of TTS voices, allowing end users and developers to adjust speech characteristics according to their preferences. Furthermore, continuous feature vectors could be particularly beneficial for synthesising speech in under-resourced languages, where traditional phonological data may be limited. By leveraging the flexibility of continuous features, researchers can create more robust TTS models that perform well even with sparse data.

As a brief aside, we note that the representation used here differs from other approaches with similar goals, for example style tokens, in that it operates on the phone level, and takes phonetic-phonologically sound assumptions as its starting point. We believe this to be an advantage or possibly a requirement if the resulting synthesis is to help shed light on fundamental phonetic-phonological questions.

Acknowledgements

This work is funded in part by the Vinnova funded project Deep learning based speech synthesis for reading aloud of lengthy and information rich texts in Swedish (2018-02427). The results will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2017-00626).

References

Acapela Group. (2005). Language manual, Swedish. http://www.acapela-vaas.com/Includes/language_manuals/Swedish.pdf

CereProc Ltd. (2023). CereVoice phone sets.

Fong, J., Taylor, J., Richmond, K., & King, S. (2019). A comparison of letters and phones as input to sequence-to-sequence models for speech synthesis. Proc. of SSW 10, 223–227. https://doi.org/10.21437/SSW.2019-40

Google. (2023). Supported phonemes and levels of stress | Cloud Text-to-Speech Documentation. Google Cloud. https://cloud.google.com/text-to-speech/docs/phonemes

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Procs. of NeurIPS 2020, 33, 17022–17033. https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html

Maniati, G., Ellinas, N., Markopoulos, K., Vamvoukakis, G., Sung, J. S., Park, H., Chalamandaris, A., & Tsiakoulis, P. (2021). Cross-lingual low resource speaker adaptation using phonological features. Interspeech 2021, 1594–1598. https://doi.org/10.21437/Interspeech.2021-327

Mehta, S., Kirkland, A., Lameris, H., Beskow, J., Székely, É., & Henter, G. E. (2023). OverFlow: Putting flows on top of neural transducers for better TTS. Interspeech 2023, 4279–4283. https://doi.org/10.21437/Interspeech.2023-1996

Mehta, S., Tu, R., Beskow, J., Székely, É., & Henter, G. E. (2024). Matcha-TTS: A fast TTS architecture with conditional flow matching. Accepted to ICASSP 2024. ICASSP 2024, Seoul, South Korea. https://doi.org/10.1109/ICASSP48485.2024.10448291

Microsoft. (2022). SSML phonetic alphabets. https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-ssml-phonetic-sets

Näslund, A., Tånnander, C., Strömbergsson, S., & Włodarczak, M. (2024). Simulating hypernasality with phonological features in Swedish TTS., Proceedings from FONETIK 2024 (pp. 81–88). Stockholm University. https://doi.org/10.5281/zenodo.11396084

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. https://doi.org/10.48550/ARXIV.2212.04356

REAPER: Robust Epoch And Pitch EstimatoR. (2023). [C++]. Google. https://github.com/google/REAPER (Original work published 2014)

Staib, M., Teh, T. H., Torresquintero, A., Mohan, D. S. R., Foglianti, L., Lenain, R., & Gao, J. (2020). Phonological features for 0-shot multilingual speech synthesis. Proc. of Interspeech 2020, 2942–2946. https://doi.org/10.21437/Interspeech.2020-1821

Swedish phonology. (2023). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Swedish_phonology&oldid=1182856646

Tånnander, C. (2018). Speech synthesis and evaluation at MTM. Proc. of Fonetik 2018, 75–80.

Tånnander, C., O'Regan, J., House, D., Edlund, J., & Beskow, J. (2024a). Prosodic characteristics of English-accented Swedish neural TTS. Proc. of Speech Prosody 2024. Speech Prosody 2024, Leiden, the Netherlands. https://doi.org/10.21437/SpeechProsody.2024-209

Tånnander, C., Mehta, S., Beskow, J., & Edlund, J. (2024b). Beyond graphemes and phonemes: Continuous phonological features in neural text-to-speech synthesis. Interspeech 2024, 2815–2819. https://doi.org/10.21437/Interspeech.2024-1565

Taylor, J., Maguer, S. L., & Richmond, K. (2021). Liaison and pronunciation learning in end-to-end text-to-speech in French. Proc. of SSW 11, 195–199. https://doi.org/10.21437/SSW.2021-34

Zandie, R., Mahoor, M. H., Madsen, J., & Emamian, E. S. (2021). RyanSpeech: A corpus for conversational text-to-speech synthesis. Interspeech 2021, 2751–2755. https://doi.org/10.21437/Interspeech.2021-341