Analysis-by-synthesis: phonetic-phonological variation in deep neural network-based text-to-speech synthesis

Christina Tånnander[1,2], David House[1] & Jens Edlund[1]

KTH Speech, Music and Hearing[1] & Swedish Agency for Accessible Media[2]
christina.tannander@mtm.se, davidh@kth.se, edlund@speech.kth.se

ABSTRACT

Text-to-speech synthesis based on deep neural networks can generate highly humanlike speech, which revitalizes the potential for analysis-by-synthesis in speech research. We propose that neural synthesis can provide evidence that a specific distinction in its transcription system represents a robust acoustic/phonetic distinction in the speech used to train the model.

We synthesized utterances with allophones in incorrect contexts and analyzed the results phonetically. Our assumption was that if we gained control over the allophonic variation in this way, it would provide strong evidence that the variation is governed robustly by the phonological context used to create the transcriptions.

Of three allophonic variations investigated, the first, which was believed to be quite robust, gave us robust control over the variation, while the other two, which are less categorical, did not afford us such control. These findings are consistent with our hypothesis and support the notion that neural TTS can be a valuable analysis-by-synthesis tool for speech research.

Keywords: analysis-by-synthesis, latent phonetic features, phonological variation, neural TTS

1. INTRODUCTION

The outstanding success of deep neural network-based text-to-speech synthesis (TTS) in recent years has not gone unnoticed (TTS usually refers to all types of text-to-speech synthesis; in the following we use the term to refer to deep neural network-based text-to-speech synthesis only). In some cases, it is indistinguishable from human speech. Even more impressive and perhaps surprising: it can be and is routinely trained on plain orthographic text. Although some findings show that phonetic or phonological transcriptions improve TTS quality (see e.g. [1, 2]), there are examples where grapheme and phoneme input yield similar results (e.g. [3]). In our own work, we have observed that a more detailed phonemic representation improves the TTS quality significantly. What is not as widely known is how the transcription (or symbol sequence) used to represent speech impacts the synthesized speech in complex ways. Previous work has studied the impact of *suprasegmental* properties, in particular how duration can act as a proxy for prominence [4, 5]. In contrast, the present work focusses on symbols representing *segmental* speech properties. In essence, this work has dual goals:

Firstly, an improved understanding of the relationship between symbolic representation in training and synthesis can lead to better TTS, which is a key goal in speech technology and one which has societal benefits, perhaps most importantly in terms of increased accessibility – a central goal in many societal policy agendas [6].

Secondly, analyzing the largely opaque but almost unreasonably successful TTS models can provide new insights. We propose, as a novel application of analysis-by-synthesis [see e.g. 7], that post hoc interpretation [8] of TTS output as a function of its input can provide information on *whether a context-dependent phonetic-phonological distinction included in the transcription is reliably realized in the speech used to train the TTS*.

In this paper, we focus on the second of these goals, and conduct initial experiments to investigate if we can indeed learn from post-hoc interpretation of TTS output. We attempt to validate the proposed analysis-by-synthesis method using three well-known phonetic-phonological allophonic variations in Swedish. The first of these variations, the lowering of /ɛ/ and /ø/ in front of /r/, is known to be categorical and in complementary distribution in the dialect represented in our experiments. The second variation, the (de)aspiration of stop consonants following /s/, is described as less robustly governed by phonological context. The third variation, the allophonic variation of /r/ itself, is expected to be more difficult to capture due to its freer nature. We trained multiple TTS models systematically varying the symbol representations of these phonemes and analyzed their outputs in relation to these variations. Our results are consistent with our expectations and support the potential usefulness of TTS as a research tool for speech science.

2. BACKGROUND

2.1. Current TTS research

Most current TTS publications focus squarely on the machine learning (ML) aspect of TTS. They are primarily concerned with the algorithms and the mathematical fit of the model and use speech primarily as a case study for testing. For reasons of comparability of the algorithms used, that type of work exhibits a strong preference for pre-processed, well-known data sets. As an example, the LJ Speech dataset for English TTS [9], with fixed transcriptions and fixed preprocessing of the audio as well as the text, is used for the vast majority of all publications on English TTS. While these studies may be useful for speech technology, they have limited value for speech science. In contrast, we look in this paper to *analysis-by-synthesis* [7] as a method to investigate the nature of speech and provide insight into phonetics and phonology.

2.2. Latent phonetic features

Modern TTS systems are typically trained on a symbol sequence representing the orthographic text and a sequence of spectrograms representing the corresponding speech. This means that no explicit information is provided on the pronunciation of words in isolation or in context, and the representation of the acoustic signal is severely impoverished. Despite this, the speech generated by systems trained in this manner can rate on par with human speech [10, 11], at least within the limitations of current evaluation metrics [12–14].

The fact that TTS systems generate speech reminiscent of a native human speaker despite purely orthographic transcriptions suggests that the models capture pronunciation implicitly. In machine learning, characteristics that are encoded implicitly and opaquely in the models are often referred to as *latent features*, and in text processing, *latent semantic features* denote implicit semantics learned by language models [15]. By analogy, we discuss *latent prosodic features* in [4], and here we use *latent phonetic features* to refer to implicitly learned segmental speech characteristics which can provide insights into phonetics and phonology.

2.3. Effects of symbol inventory in neural TTS

TTS symbol sets that go beyond the graphemes often include representations for a range of speech phenomena. Most commonly, they hold phones or phonemes, stress, and boundary types. The symbol sets of four companies developing Swedish TTS [16–19] all include separate symbols for the lowered variants of short and long /ɛ/ and /ø/. One included separate symbols for aspirated /p, t, k/. None included more than one symbol for Swedish /r/, although some used separate symbols for non-Swedish /r/ sounds (e.g. English [ɹ]).

The selection of speech phenomena in the TTS symbol inventory impacts the TTS output. [20] trained several systems on podcast speech with different symbolic representations of filled pauses. Omitting filled pauses in the transcriptions resulted in a system that inserted filled pauses in reasonable places, but without the means to prevent these insertions. Using a single token for “uh” and “um” allowed for the removal of filled pauses and their insertion in specific places, but when the symbol was used in the text, the choice of "uh" or "um" was made implicitly by the system. Transcribing "uh" and "um" as separate tokens gave the user explicit control over both placement and type. We have observed the same effect with other phenomena, such as audible breath, in our own work. It may be worth noting in this context that explicit control is not always desirable in TTS. While the ability to control filled pauses allows the user to specify their placement, a TTS that inserts them in appropriate places on its own can solve this task without requiring user knowledge. In short, the two solutions will be good for different tasks.

We aim to use the insights from [21] to investigate segmental properties of a single Swedish speaker's speech. Specifically, we will examine three phonetic-phonological variations in Swedish.

2.4. Allophones of the vowels /ɛ/ and /ø/

In many Swedish dialects, and particularly in the eastern part of Sweden including the Stockholm area, the long vowels /ɛː/ and /øː/ are pronounced as the more open allophones [æː] and [œ̝ː] when preceding /r/ or a retroflex consonant [22–25]. This also applies to the short vowels /ɛ/ and /œ/ which are lowered to [æ] and [ɶ] in the same context. The lowering of short vowels is even more widespread among Swedish dialects than for long vowels [22]. There are exceptions to this pronunciation, such as when a morpheme boundary occurs after the vowel [23]. This variation is known as a robust complementary distribution and is clearly present in our training data.

2.5. Aspiration in the stop consonants /p/, /t/ and /k/

One of the most widely referred to phonetic variations used to illustrate complementary distribution in Swedish (as well as English) is the aspiration of the voiceless stops /p/, /t/ and /k/ and the absence of this aspiration when the stop is preceded by an /s/ in the same syllable [23, 24, 26, 27]. Following a morpheme boundary, such as the initial position of the second

Model	Context for additional symbols	Original symbols (N)	Additional symbols (N)
TTS$_{\text{ÄÖ}}$	Vowels before /r/ or retroflex	/ɛː/ (2073) /øː/ (2013) /ɛ/ (47996) /œ/ (1797)	[æː] (2563) [œː] (2763) [æ] (2515) [ɶ] (1981)
TTS$_{\text{PTK}}$	Voiceless stops after /s/ in same syllable	/p/ (9394) /t/ (36939) /k/ (17737)	[p] (723) [t] (4251) [k] (2424)
TTS$_{\text{R}}$	/r/ following voiceless stop in same syllable	/r/ (46403)	[r] (1841)

Table 1. Composition of the three models with separate symbols for pairs of allophones. In the base model **TTS$_{\text{R}}$** the original symbol is used for all allophones.

component of a compound (e.g. "*ås-topp*", en. "ridge") the stop will generally retain its aspiration [22, 23]. The strength of aspiration can vary depending on many factors, and the strongest aspiration generally occurs in word-initial position in stressed syllables. Aspiration is often weaker and can even disappear in unstressed syllables [22, 24].

2.6. Free variation of /r/

The phoneme /r/ exhibits particularly rich phonetic variation and is often referred to as *the class of rhotics*. This variation occurs not only between languages, but also within languages representing both dialectal variation and individual speaker variation. Phonetic realization of /r/ includes trills, taps, flaps, fricatives and approximates as well as the influence on vowel quality by a following /r/. In addition to this variation in manner of articulation, place of articulation can range from dental to uvular [24, 28–30]. In Swedish, we find a pronounced dialectal variation in place of articulation. Southern Swedish dialects employ uvular place of articulation while central and northern Swedish dialects use alveolar place. In between, we find large areas where both places are found in complementary distribution [26, 31, 32].

The rhotics of the central Swedish dialect and the Stockholm area depends on individual preferences, the articulation force used and the degree of formality [24, 31]. However, both Elert and Malmberg point out that trills and taps are more common following a syllable-initial consonant (e.g. *prova, träna, kråka*). [31] includes syllable-initial /r/ before a stressed vowel as a position favoring a trill in certain situations but a fricative realization is also common. This variation is noted to be partially regulated by vocal effort and speaker preferences. According to [31], the fricative is the most common realization following a vowel (e.g. *bara*) or in word-final position (e.g. *pojkar*). [23] even lists /r/-deletion in unstressed positions as characteristic of the Stockholm dialect.

3. METHOD

Our proposed method starts with a TTS model that can produce a specific allophonic variation represented by a single symbol **S**. We then train a new model on the same speech data, but with an extended symbol inventory that separates allophones into separate symbols **S$_1$** and **S$_2$** based on phonological context rules. We test the model by synthesizing an utterance with **S$_1$** in a context where **S$_2$** is expected, or vice versa. If the mismatched allophone is noticeable to human listeners, we have gained explicit control over the allophones, otherwise we have not.

3.1 Assumptions and hypothesis

We assume that if we gain explicit control over the allophones in this procedure, it is because the phonological distinction expressed by the rule matches the distribution of acoustic/phonetic variation in the speech data, which is consistent with the variation being governed by the phonological distinction. Failure to control the allophones may occur for a wide range of reasons and is not informative.

We hypothesize that we will be able to control the famously robust and categorical /ɛː, øː, ɛ, œ/ allophones, and unable to control the notoriously free /r/ variation. As for aspiration of consonant stops, we defer prediction as this variation is described as robust in some contexts and much less so in others.

3.2 Stimuli design

We designed stimuli to test our hypothesis using four Swedish neural TTS voices trained on the same audio data of a professional female speaker from the Stockholm area. **TTS$_{\text{BASE}}$** used a minimal phonemic symbol inventory with extra symbols for pauses and word boundaries. The other three models (**TTSÄÖ**, **TTSPTK** and **TTSR**) added explicit symbols for our selected allophonic pairs of /ɛː, øː, ɛ, œ/, /p, t, k/ and /r/ to the symbol inventories (see Table 1). Each model was trained for 600 epochs (iterations of the entire training data) in Nvidia's PyTorch implementation of Tacotron 2 [33] and synthesized with the WaveGlow vocoder [34] trained 650 epochs on the same voice as the Tacotron models.

We designed three sets of test sentences to showcase each of the three Swedish phonological variations. All sentences were synthesized with the **TTS$_{\text{BASE}}$** model. Next, each set of test sentences was synthesized with its respective TTS model in two ways: (1) with the specific symbols in their correct places according to the context and (2) with the opposite symbol of what the context would dictate. We synthesized each sentence five times in each condition to provide a sense of the variation in the models.

3.3 Analyses

To evaluate our hypotheses, we conducted the following analyses: (1) We evaluated the general capability of the **TTS$_{\text{BASE}}$** model by listening to and comparing the synthesized sentences to the original speaker's voice, dialect, and speaking style. (2) Three experienced Swedish phoneticians performed a qualitative analysis of the renditions with interchanged allophones for each series of test sentences to determine if our expectations were met, and our hypothesis supported. (3) For the /p, t, k/ test sets, we also conducted a quantitative analysis of voice onset time (VOT), a strong predictor of aspiration.

4. RESULTS

The base model performed well for all test sets. Even though 600 is a quite low number of epochs when training Tacotron 2, the TTS was similar to the original speaker in terms of voice, dialect, and speaking style. No major differences were observed between the test sets, and the allophones of interest were all produced in an appropriate manner.

For **TTSÄÖ**, the renditions with the allophone symbols correctly placed were indistinguishable from the base model. In contrast, interchanging the symbols consistently produced the allophones in a manner that sounded like proper speech, but highly unlikely for the specific voice due to the context. In pre-rhotic position, the more closed allophones gave a clear impression of a different dialect, and in non-pre-rhotic position, judgements varied between yet another dialect or, if semantics permitted, the same word with an infixed /r/, such that "*höna*" ([hœ̞ːna], en. "hen") was perceived as "*hörna*" ([hœ̞ːɳa], en. "corner").

For **TTSPTK**, interchanging the symbols used to capture aspirated and non-aspirated voiceless stops did not produce any clear perceptual differences in the renditions. As a follow-up, we measured voice onset time (VOT) of stops after /s/ in the same syllable. We omit the details here since again, no differences were found.

Finally, for the **TTSR**, models, the general quality was indistinguishable from the base model regardless of whether the symbols were interchanged. Two listeners perceived that a more pronounced, sometimes trilled /r/ was used more often when the symbol intended to capture trills was used, but we were unable to quantify this observation. The model produced a relatively wide range of /r/ realizations, all sounding appropriate, but these variations could not be predicted by phonological context.

5. DISCUSSION

Our primary goal was not to experiment with phonetic control for TTS production purposes, but rather to use TTS as a tool to gain insights into the powerful latent phonetic features in neural TTS systems. Our hypothesis that using separate symbols for /ɛː, øː, ɛ, œ/ and their lowered allophones, which are reliably governed by phonological context, would allow us to control the distribution of these allophones was borne out. This supports the notion that the ability to control a variation using different symbols that are assigned based on a principle (such as phonological context rules) indicates the accuracy of that principle in describing the variation. By utilizing the complex but inaccessible TTS models as a tool in conjunction with our perception, we can study speech beyond the limitations of easily measurable and accessible phonetic features. While this is currently a proof-of-concept, it has the potential to expand our understanding of speech.

6. 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).

7. REFERENCES

[1] 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*, ISCA, 223–227.

[2] 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*, ISCA, 195–199.

[3] Perquin, A., Cooper, E., Yamagishi, J. 2020. *Grapheme or phoneme? An analysis of Tacotron's embedded representations*.

[4] Tånnander, C., House, D., Edlund, J. 2022. Syllable duration as a proxy to latent prosodic features. *Speech Prosody 2022*, ISCA, 220–224.

[5] Tånnander, C., Edlund, J. 2021. Stress manipulation in text-to-speech synthesis using speaking rate categories. *Proc. of Fonetik 2021*, Lund University, 17–22.

[6] United Nations. 2015. *Transforming our world: the 2030 agenda for sustainable development*. Department of Economic and Social Affairs, United Nations.

[7] Bever, T. G., Poeppel, D. 2010. Analysis by synthesis: a (re-)emerging program of research for language and vision. *BIOLINGUISTICS* 4, 2–3: 174–200.

[8] Lipton, Z. C. 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3: 31–57.

[9] Ito, K., Johnson, L. 2017. *The LJ Speech dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[10] Shen, J., Pang, R., Weiss, R. J., et al. 2018. Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions. *Proc. of ICASSP 2018*, 4779–4783.

[11] Tan, X., Qin, T., Soong, F., Liu, T.-Y. 2021. *A survey* on neural speech synthesis. arXiv 2106.15561, Jun. 2021.

[12] Wagner, P., Beskow, J., Betz, S., et al. 2019. Speech synthesis evaluation - state-of-the-art assessment and suggestion for a novel research program. *Procs. of SSW10*, ISCA, 105–110.

[13] King, S. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1.

[14] Shirali-Shahreza, S., Penn, G. 2018. MOS naturalness and the quest for human-like speech. *In procs. of SLT 2018*, 346–352.

[15] Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., Miliani, M. 2022. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*.

[16] Acapela Group. 2005. Language manual, Swedish. [Online]. Available: http://www.acapela-vaas.com/Includes/language_manuals/Swedish.pdf

[17] Amazon. 2022. Amazon Polly: Developer guide. [Online]. Available: https://docs.aws.amazon.com/pdfs/polly/latest/dg/polly-dg.pdf

[18] CereProc. 2022. CereVoice phone sets. [Online]. Available: https://www.cereproc.com/files/CereVoicePhoneSets.pdf

[19] Microsoft. 2022. SSML phonetic alphabets. [Online]. Available: https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-ssml-phonetic-sets

[20] Székely, É., Henter, G. E., Beskow, J., Gustafson, J. 2019. Spontaneous conversational speech synthesis from found data. *Interspeech 2019*, ISCA, 4435–4439.

[21] Székely, É., Eje Henter, G., Beskow, J., Gustafson, J. 2019. How to train your fillers: uh and um in spontaneous speech synthesis. *Proc. of SSW 10*, ISCA, 245–250.

[22] Bruce, G. 2010. *Vår fonetiska geografi - Om svenskans accenter, melodi och uttal*. Studentlitteratur, Lund.

[23] Elert, C.-C. 1970. *Ljud och ord i svenskan*. Almqvist & Wiksell Förlag AB, Stockholm.

[24] Malmberg, B. 1968. *Svensk fonetik*. Gleerups, Lund.

[25] Schumacher, A. 2014. Hät, här, höt and hör - An articulatory and acoustic study of the Swedish vowels /ɛː/ and /øː/, Master Thesis, Lund University, Lund.

[26] Elert, C.-C. 1981. *Ljud och ord i svenskan 2*. Almqvist & Wiksell International, Stockholm.

[27] Ladefoged, P. 1982. *A course in phonetics*. Harcourt Brace Jovanovich, Inc., New York.

[28] Ladefoged, P., Maddieson, I. 1996. *The sounds of the world's languages*. Blackwell Publishers, Oxford.

[29] Laver, J. 1994. *Principles of phonetics*. Cambridge University Press, Cambridge.

[30] Lindau, M. 1980. *The story of /r/*. UCLA.

[31] Elert, C.-C. 1995. *Allmän och svensk fonetik*. Norstedts Förlag AB, Stockholm.

[32] Sjöstedt, G. 1936. Studier över r-ljuden i sydskandinaviska mål.

[33] Wang, Y., Skerry-Ryan, R. J., Stanton, D., et al. 2017. Tacotron: towards end-to-end speech synthesis. *Proc. of Interspeech 2017*, ISCA, 4006–4010.

[34] Prenger, R., Valle, R., Catanzaro, B. 2019. Waveglow: a flow-based generative network for speech synthesis. *Proc. of ICASSP 2019*, IEEE, 3617–3621.