Stress manipulation in text-to-speech synthesis using speaking rate categories

Christina Tånnander[1,2] and Jens Edlund[1]
[1]Speech, Music and Hearing, KTH
[2]Swedish Agency for Accessible Media, MTM

Abstract

The challenge of controlling prosody in text-to-speech systems (TTS) is as old as TTS itself. The problem is not just to know what the desired stress or intonation patterns are, nor is it limited to knowing how to control specific speech parameters (e.g. durations, amplitude and fundamental frequency). We also need to know the precise speech parameters settings that correspond to a certain stress or intonation pattern – over entire utterances.

We propose that the powerful TTS models afforded by deep neural networks (DNN's), combined with the fact that speech parameters often are correlated and vary in orchestration, allow us to solve at least some stress and intonation parts by influencing a single easy-to-control parameter, rather than detailed control over many parameters.

The paper presents a straightforward method of guiding word durations without recording training material especially for this purpose. The resulting TTS engine is used to produce sentences containing Swedish words that are unstressed in their most common function, but stressed in another common function. The sentences are designed so that it is clear to a listener that the second function is the intended. In these cases, TTS engines often fail and produce an unstressed version.

A group of 20 listeners compared samples that the TTS produced without guidance with samples where it was instructed to slow down the test words. The listeners almost unanimously preferred the latter version. This supports the notion that due to the orchestrated variation of speech characteristics and the strength of modern DNN models, we can provide prosodic guidance to DNN-based TTS systems without having to control every characteristic in detail.

Introduction

If we diminish speech and speaking to a simple one-way information transfer task, stress and intonation may not be as crucial as other speech features. No matter how hard we try to showcase Swedish minimal pairs such as “malen” (noun; Eng. “the moth”; Swedish accent I) and “malen” (past participle, Eng. “ground”; Swedish accent II), in real spoken communication, these are unlikely to cause actual misunderstandings. If intonation and stress were necessary for understanding, then we would not be able to make any sense at all of these minimal pairs in writing. But, in both writing and speech, the context is often sufficient to grasp the intended meaning, unusual intonation patterns notwithstanding.

A more interesting question is how we *perceive* speech with atypical or unexpected stress and intonation. Setting any discussion of dialects and native speaker variation aside, we find atypical stress and intonation in the speech of second language speakers, where it gives an accurate impression of what it is: accented speech. We also find it in text-to-speech synthesis systems (TTS), where it may be a result of text processing that is unaware of the relation between words, or a sound generation system that does not afford control of stress and intonation. Until recently, the general quality of TTS was poor enough that listeners would be implicitly aware that they were listening to a machine rather than a human, and the interpretation of misplaced stress and/or atypical intonation patterns would likely involve this fact. However, modern, DNN-based TTS such as Tacotron (Wang et al., 2017), is hard to tell apart from human speakers, and listeners may face a situation where they hear a perfectly human and native-sounding speaker make unusual intonation and stress choices.

Under these circumstances, misunderstandings may actually occur, but more importantly, the cognitive load involved in listening likely goes up as a result of these surprises. We may safely assume that listeners prefer and are helped by typical, expected stress and intonation patterns.

In this paper, we describe a way of controlling speaking rate by simple means in a DNN TTS, and propose that by controlling per-word speaking rate, we can target the desired stress and intonation patterns. We support this with a limited study in which we target the realisation of two well-known types of homograph pairs in Swedish: stressed verb particles/unstressed prepositions and stressed numerals/unstressed indefinite articles, assuming that the detection of these parts of speech is already solved, and believing that the same method can be applicable also on other phenomena where prominence is involved, such as contrastive focus.

Background

Spoken text

We focus here on *spoken text* (Tånnander & Edlund, 2019), and more specifically text that is read aloud with the assistance of a TTS voice. We assume a text preprocessing step that accurately analyses the structure of the text to be read, so that we can tell for example particle verbs in phrasal verbs apart from other verb + preposition constructs, and that we can tell the indefinite articles “en” and “ett” (Eng. “a/an”) apart from the numerals “en” and “ett” for number 1 (Eng. “one”) (these are homographs in Swedish).

The main driver for the work is practical – we want to be able to read texts aloud with TTS in such a manner that it can be understood with a minimum of cognitive load. The end goal is to increase the accessibility of Swedish texts by improving talking books that for a variety of reasons must be produced with TTS.

Talking books

Talking books are produced by the Swedish Agency for Accessible Media (MTM) for people with vision impairments or reading difficulties in accordance with Section 17 of the Swedish Copyright Act. The target groups studying at university level depend on a fast production of talking textbooks, and MTM produce around 1 000 Swedish and English talking books per year with TTS, as well as more than 100 newspapers (Tånnander, 2018).

Intonation and stress in TTS

Modern DNN-based TTS is trained on large amounts of data and often sound very human-like. The speech it produces reflects the prosodic features of the training data well, but may come out a bit bland, as it tends towards averages in the training data (Wang et al., 2018) Even if the desired prosodic patterns can be assigned a text in theory, the realisation of prosodic features is limited to what can be controlled in the TTS. And whereas more traditional TTS methods (e.g. formant and concatenative TTS) are to some degree transparent and allow for trouble shooting and manual correction of problematic utterances, the inner workings of the DNN are more obscure. This has radically changed the landscape of research into how speech characteristics in TTS can be controlled.

Examples of systems that control speech characteristics include Shechtman & Sorin (2019), who showed that they could control expressiveness and per-sentence speaking rate without losing quality and naturalness. Raitio et al. (2020) included pitch, pitch range, phone durations, energy and spectral tilt in their model, making it possible to control these features per utterance, independently of each other. The evaluation showed a significant decrease in MOS score when slowing down or speeding up the speech. There is a growing number of similar efforts around, but to date, we are not aware of any Swedish DNN-based TTS system that manages sets and generates the appropriate stress and intonation in spoken text.

Method

Assumptions and hypothesis

Assumption 1: Speech characteristics commonly associated with prominence, emphasis and stress include increased phoneme durations, higher fundamental frequency and increased amplitude, and these are correlated. We assume, that at least in some contexts, variations in fundamental frequency and amplitude can to some extent be predicted from variations in phoneme durations.

Assumption 2: DNN-based TTS learns, given a character string (e.g. an orthographic, phonemic or phonetic transcription), to produce the acoustic sequence (in actuality a spectrogram, but for our purposes, we will view it as a sound signal) that is maximally typical for what it has previously associated with similar strings.

Assumption 3: Intonation patterns are influenced by phrase stress, and to some extent predictable given that we know where it is located.

We hypothesise, then, that if we can influence syllable durations in DNN-based TTS for constituents that commonly occur in both stressed and unstressed positions, and with different meanings in the training material, we will trigger the TTS to choose a realisation that is appropriate for the stressed variety when we force a longer duration, and for the unstressed variety when we force a shorter duration. In these positions (at least), fundamental frequency and amplitude will follow duration to create realistic stress realisations, and the rest of the utterance will in turn be given a realistic intonation pattern given that stress. In short, we assume that *in DNN-based TTS, per-word control over duration is sufficient to control intonation patterns that are associated with phrase stress*

Training data

Almost 18 hours of a female professional voice was used as training data. The voice was originally recorded for a unit selection voice for non-fiction texts (Tånnander, 2018) and the voice talent was told to read the prompts in a clear and controlled manner, without too much prosodic variation. The training data was represented as phonemic transcriptions, including the Swedish stress markers (accent I, accent II, and secondary stress). Pauses within sentences were marked with '/' and word boundaries with '&'.

The data had been phonemically segmented in the preparation process for the unit selection voice, giving us access to the start and end points for each phoneme. The distribution of syllable lengths is shown in Figure 2. The average syllable duration was 247 ms.

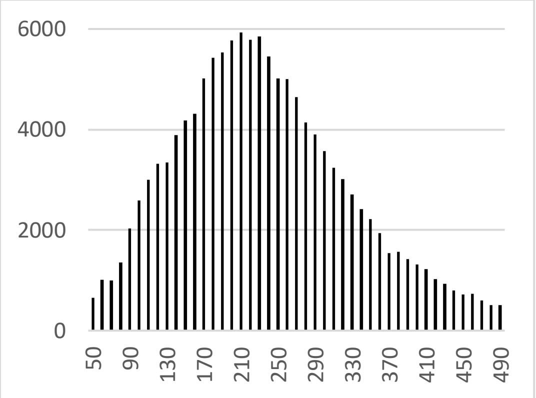

Figure 2. Syllable length (ms) distribution (x axis) and frequencies (y axis) in the training data. Syllable lengths occurring less than 500 times are not shown.

Preprocessing of the training data

All words in the training data were categorized into one of five approximately equal-sized speaking rate categories with around 27 000 words in each, as shown in Table 1.

Table 1. The five speaking rate categories. (1) is the fastest and (5) the slowest speaking rate.

Speaking rate category	Average syllable duration in word (ms)	Occurrences
1	< 150	27 152
2	151- 210	26 075
3	211- 250	28 042
4	251- 315	27 509
5	> 316	27 010

A label for the corresponding group was added to the word tokens in the training data. However, adding extra control features to the metadata makes it necessary to provide the same control features in the input to the TTS system. Here, we only want to control the durations of some of the words in the input, not all of them, so the model must support tokens without durational information as well. As a general fix that does not duplicate data or increase the size of the model, we added the speaking rate categories only to every second utterance in the training data, which in effect trains a model without durational data on half the training material. For the other half, the speaking rate category (1-5) was simply inserted before each word in the metadata with the phonemic representations of the speech. This allows us to guide the speaking rate not only of entire sentences, but of single words.

Note that this change requires no changes to the DNN structure or training framework, it all takes place in the preprocessing of the training data.

TTS training

The voice was trained in a PyTorch framework, and ran 1 500 epochs of decoder training (wav to mel spectrogram) with Tacotron (Wang et al., 2017) and 650 epochs of encoder training (mel spectrogram to wav) with the WaveGlow vocoder (Prenger et al., 2019).

Test data and setup

The test data contained 20 sentences, 10 with a verb particle that could potentially be confused with an unstressed preposition or another function word (see the example sentence in Table 2, and 10 with a numeral (*en* or *ett*, see Table 3), which could potentially be confused with its homographic counterpart in Swedish, the unstressed indefinite article.

Table 2. Example of a test sentence with a phrasal verb with a verb particle.

Viktor	tänker	hoppa av	gymnasiet.
Victor	intends to	drop out of	high school.

Table 3. Example of a test sentence with a numeral.

Ett	träd	har	fler	än	ett	löv.
A	tree	has	more	than	one	leaf.

All test sentences are shown in Table 4. Note that sentences where both readings would make sense, but mean different things, were avoided. The goal of this experiment was not to prove that ambiguity exists (in rare cases), and that prosody can be used to disambiguate in these rare cases. We already know this, and although it may be exciting from a research point of view, its value in real-world situations is small. Rather, we want to show that we can create more acceptable readings by guiding the TTS towards the expected prosody using simple means.

Each sentence was synthesized twice, once without any speaking rate categories (**Voice1**), and once with the speaking rate category 5 before the verb particle or numeral, to signal that the test word should be pronounced in '*slow mode*', that is with longer durations (**Voice2**). The verbs in the phrasal verb constructions were marked with the speaking rate category 2, to signal that the verb should be destressed. In this way, we hoped to guide the TTS towards a reading in which these words were made more generally more prominent, and be perceived as stressed, with the broader effect that the entire sentence melody was adapted to such a reading.

Experiment

20 native Swedish listeners participated, all employees at the Swedish Agency for Accessible Media, MTM.

A web-based *Comparison Category Rating* (CCR) test was used, in which participants listened to each sentence pair and marked one of them as preferred. The order of the sentence pairs was randomized, as was the order of the **Voice1** and **Voice2** within each pair.

*Table 4. The stimuli type, preferences for **Voice1** and **Voice2**, proportional preferences for **Voice2**, and the sentence with the stressed verb particle or numeral in bold.*

Stimulus	Type	Voice1	Voice2	%	Sentence
1	VP	3	17	85%	Staten lyckades till slut komma **på** ett sätt att delta i finansieringen.
2	VP	2	18	90%	Jag tänker köra **på** så länge det lönar sig .
3	VP	2	18	90%	Alla visste **om** det utom jag.
4	VP	1	19	95%	Lisa steg **in** i salen.
5	VP	7	13	65%	Pelle gick **fram** och tog trofén.
6	VP	6	14	70%	Karina ska hälsa **på** sin guddotter på söndag.
7	VP	5	15	75%	Vet du **om** att det är torsdag imorgon.
8	VP	5	15	75%	Det kan hända att du inte tycker **om** det.
9	VP	1	19	95%	Pernilla måste lägga **av** med oförskämdheterna.
10	NUM	0	20	100%	Viktor tänker hoppa **av** gymnasiet.
11	NUM	0	20	100%	Allt städades med undantag av **ett** rum.
12	NUM	4	16	80%	Rickard åt bara **en** enda kaka, sen var han nöjd.
13	NUM	0	20	100%	Ett träd har inte bara **ett** löv.
14	NUM	9	11	55%	**Ett** av borden på gräsmattan är rangligt.
15	NUM	2	18	90%	Det är **en** sida av saken.
16	NUM	3	17	85%	Pojken fyller **ett** om två veckor.
17	NUM	0	20	100%	Det var bara **en** person som anmälde sig.
18	NUM	3	17	85%	Ett plus **ett** är lika med två.
19	NUM	15	5	25%	Fyra hästar, tre hönor, **en** kossa och sju grisar.
20	NUM	0	20	100%	Femte paragrafen stycke **ett** till tre.

Results

The readings in which the test word (a verb particle or numeral) was guided towards a slower pace (Voice2) were preferred 332 of 400 times (83%: 82% for numerals and 84% for verb particles). The result for each stimulus is presented in Table 4.

All 20 individual participants preferred Voice2, and the participant that chose Voice1 the most times did so in only 7 of 20 comparisons. The participants were free to listen to the stimuli pairs in any order, and were also allowed to listen to them as many times they wanted to, so effects of order are not well controlled for. Regardless, the voice presented last was preferred in 57% of the cases. Voice2 scored higher than Voice1 in all individual stimuli pairs except one, where Voice1 was preferred 15 of 20 times (stimulus 19). In one other pairwise comparison the result was close: 45% preferred Voice1 and 55% Voice2.

Discussion

The results are straightforward: participants preferred Voice2, and there was not much variation among users or stimuli. Only one stimuli pair showed preferences for Voice1 (75%), see stimuli 19 in Table 4. This can be explained by the fact that the sentence (*"Four horses, three hens,* ***one/a*** *cow and seven pigs."*) was a mistake on our behalf: it actually works well without treating *'en'* as a numeral. The TTS simply did a better job with that reading of the sentence.

Another stimulus that stands out is number 14, where 55 % preferred Voice2 (*"****One*** *of the tables on the lawn is rickety."*). Here, the numeral *'ett'* is sentence initial and followed by a preposition, which gives a hint that 'ett' should be stressed. The Tacotron may just have learned this example from the training data without guidance.

Apart from that, the result is as expected: without guidance, the system produces the statistically most common reading of the test words: unstressed function words and indefinite articles. This produces perfectly comprehensible readings, but with an intonation that is unexpected. With a nudge towards a reading where the test words are lengthened, the system produces an utterance that is more aligned, as a whole, to that lengthening and therefore preferrable to the listeners.

Conclusions and future work

The results show that it is possible to create TTS readings of Swedish verb particles and numerals that are homographs to various function words and indefinite articles, by guiding the TTS model towards a lengthening of these words, and that the resulting overall reading is preferrable to one where no such guidance was provided. We argue that without the lengthening, the more common unstressed reading is likely to be chosen, while the lengthening causes the models to choose a path that is common for slower readings of these words: the more prominent verb particles and numerals.

It should be noted that the differences in stress level for the verb particles and numerals are not isolated. In fact we hope that the proposed method is a viable way to provide prosodic guidance to a DNN-based TTS system without having to control every parameter in detail (something we simply do not know how to do, in part since we do not know what the correct targets are). We believe that the method can be used to set word prominence in other situations, such as emphasising new information.

Acknowledgments

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

Prenger, R, Valle, R, & Catanzaro, B (2019). Waveglow: A flow-based generative network for speech synthesis. Proc. of ICASSP 2019, 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143

Raitio, T, Rasipuram, R, & Castellani, D (2020). Controllable neural text-to-speech synthesis using intuitive prosodic features. Proc. of Interspeech 2020, 4432–4436. https://doi.org/10.21437/Interspeech.2020-2861

Shechtman, S, & Sorin, A (2019). Sequence to sequence neural speech synthesis with prosody modification capabilities. Proc. of the 10th ISCA Speech Synthesis Workshop, 275–280. https://doi.org/10.21437/SSW.2019-49

Tånnander, C (2018). Speech Synthesis and evaluation at MTM. Proceedings of Fonetik, 75–80. https://www.ling.su.se/fonetik-2014/tidigare-konferenser

Tånnander, C, & Edlund, J (2019). First steps towards text profiling for speech synthesis. Proc. Digital Humanities in the Nordic Countries 2019 (DHN2019).

Wang, Y, Skerry-Ryan, R J, Stanton, D, Wu, Y, Weiss, R J, Jaitly, N, Yang, Z, Xiao, Y, Chen, Z, Bengio, S, Le, Q, Agiomyrgiannakis, Y, Clark, R, & Saurous, R A (2017b). Tacotron: Towards end-to-end speech synthesis. Proc. of Interspeech 2017, 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452

Wang, Y, Stanton, D, Zhang, Y, Skerry-Ryan, R J, Battenberg, E, Shor, J, Xiao, Y, Ren, F, Jia, Y, & Saurous, R A (2018). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. International Conference on Machine Learning, 2018. https://arxiv.org/abs/1803.09017