11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Methods of slowing down speech

Christina Tånnander[1,2], *Jens Edlund*[1]

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

Abstract

A slower speaking rate of human or synthetic speech is often requested by for example language learners or people with aphasia or dementia. Slow speech produced by human speakers typically contain a larger number of pauses, and both pauses and speech have longer segment durations than speech produced at a standard or fast speaking rate.

This paper presents several methods of prolonging speech. Two speech chunks of about 30 seconds each, read by a professional voice talent at a very slow speaking rate, were used as reference. Seven pairs of stimuli containing the same word sequences were produced, one by the same professional, reading at her standard speaking rate and six by a moderately slow synthetic voice trained on the same human voice. Different combinations of pause insertions and stretching were used to match the total length of the corresponding reference stimulus. Stretching was applied in different proportions to speech and non-speech, and pauses were inserted at punctuations, at certain phrase boundaries, between each word, or by copying the pause locations of the reference reading.

128 crowdsourced listeners evaluated the 16 stimuli. The results show that all manipulated readings are less consistent with expectations of slow speech than the reference, but that the synthesised readings are comparable to stretched human speech. Key factors are the relation between speech and silence and the duration of talkspurts.

1. Introduction

Language learners and people with cognitive impairments (e.g. aphasia or dementia) often prefer a slower speaking rate when listening to longer texts read aloud. A number of studies have attempted to find a balance between speaking rate and comprehension among aphasics (see for example [2]–[4]).

The most obvious method to produce read speech with a slow speaking rate is to instruct a voice talent to read very slowly, but to create books or other long texts at different speaking rates with human readings would be prohibitively expensive. Using slow speech synthesis trained on or adapted to very slow speech materials is another option, but again, this may not always be practicable.

In this study, we investigate several methods to create very slow speech using existing speech materials. The results are evaluated against human reference readings at a very slow speaking rate by a professional voice talent, and a stretched human reading in a listening experiment with 128 crowd sourced listeners.

2. Background

Talking books

We are mainly concerned, here, with texts read aloud for people with vision impairments or reading difficulties: societal information, news, and so-called *talking books*. The difference between an audiobook and a talking book varies somewhat in different countries, but is often present and similar in meaning. The Swedish Agency for Accessible Media (MTM) states that a talking book "is intended for persons with a permanent or temporary print disability", that they are "produced with public funds and in accordance with Section 17 of the Copyright Act", and that the "the recording of a talking book must conform with the original, which must be a published work" [1]. In other words, the option to simplify or otherwise change the written text to make it easier to understand is available.

Speaking rate

The speed at which speech is produced can be measured in different ways. *Speaking rate* is defined as the rate at which a certain idealised (e.g. phonological or orthographic) unit is produced per total speech and non-speech (silences, breath etc.) duration [5]. *Articulation rate*, on the other hand, is the number of actual speech units (e.g. phonemes) divided by the duration of the actual speech, non-speech such as silences and breathings excluded [6]. In speech science, *speech rate* is sometimes used with the same meaning as speaking rate, and sometimes as a metric more closely tied to the acoustic signal and its variations.

Both speaking and articulation rate can be measured in different ways, for example the number of syllables per time unit, such as syllables per second [5], [7], [8], syllables per minute [9] or average syllable duration [10]. Another common metric of speaking rate is words per minute (wpm). Since word length differ between languages and speaking situations, wpm can be a too rough metric in many situations. In research, wpm is often presented alongside metrics that reflect the pronunciation of the words and data about average syllables per word [9].

Measuring speaking rate is not trivial and even if researchers use the same metrics, the counting of words, syllables or phones can differ. There is no obvious unified way to count phones or other speech units, for example glottal stops, affricates, diphthongs or syllabic consonants [11]. Also, the number of phones or syllables can be differentiated into the *intended* number of speech units and the number of units that are actually *realized*. It has been shown that listener's perception of speaking rate reflects both the intended and realized speaking rates [12].

10.21437/SSW.2021-8

Slow speech

Slow speech is characterised by a larger number of pauses, longer pause durations and longer phone durations [5]. [6] found that the articulation rate makes up only a small part of the changes in speaking rate, and the largest change was the total pause durations. Other things affect speaking rate as well. There is evidence of regional variations in articulation rate [7], [13], and the number of pauses inserted in read speech can depend on text genre (e.g. news reports and novels) [14]. Pauses tend to be longer the more syllables there are in the utterance [14]–[16].

The purpose of a recording can also be seen in speech characteristics. TTS recordings, for example, have been characterized as having low speaking rate as well as low mean pitch and standard deviation of energy [17], and spontaneous speech as faster than read speech, with a greater variance [11].

Simply stretching speech while maintaining F_0 and other characteristics is, unsurprisingly, not consistent with human speech. In humans, a slower speaking rate correlates with a lower F_0 [18], hyperarticulated speech is characterized by a slower speaking rate, a higher number of pauses, more syllables, which altogether result in a longer total duration of speech and non-speech [8]. Perceived speaking rate is also affected by non-durational characteristics: [19] found that high, fairly monotonous speech segments lead to a higher perceived speaking rate.

Typical speaking rate

In British English, the speaking rate vary between 140 (lecture) to 210 wpm (conversation), with corresponding syllables per second of 190 and 260 [9]. Similarly, a summary of different acoustic features among different English speech corpora shows that the lowest speaking rate was found in audiobooks, followed by recordings for TTS and broadcast news, while corpora consisting of conversational speech show a higher speaking rate [17]. Proficiency matters, too. A study investigating pausing among English language learners reported that native speakers pause 7.15 times per hundred words (phw), while the learners pause much more frequently, between 10.76 to 14.43 phw, depending on proficiency.

[20] reported that a Swedish professional speaker had a speaking rate of 130 wpm in normal mode, 111 in slow mode and 106 wpm in distinctive mode (146 in fast mode). These variations were mainly associated with total pause durations (longer pauses and a larger number of pauses). At a slow speaking rate, the sum of the pause durations was almost 50% longer than at a normal speaking rate, while the phoneme durations differed only by 4%.

Controlling speaking rate in speech synthesis

Modern, unsupervised methods for training speech synthesis often capture prosody well. It does so behind the scenes, leaving limited room for investigation or control for the researcher. Control of prosody, or the lack thereof, is a well-known issue and an active research area, and in some cases, the investigation of prosody is the very reason for creating a synthetic voice. [21] controlled expressiveness and sentence wise speaking rate without losing quality and naturalness. [22] facilitated the independent control of pitch, pitch range, phone durations, energy and spectral tilt by including these in their model, but their evaluation showed a significant decrease in MOS score when slowing down or speeding up the voice. This may have been a result of an overly generic evaluation question, confounding for example a dispreference for slow speech with a poor quality rating for slow speech.

3. Method

Participants

Listeners were recruited through Prolific, a subject pool for online experiments [23]. At the time of the experiment, Prolific had 815 active subjects between the ages of 18 and 67 reporting as fluent in Swedish. We recruited 64 of these for each of two utterances, totalling 128 sessions, and paid marginally above the recommended fee. Each test took between 5 and 6 minutes to complete and listeners were rewarded £0.8. Listeners were allowed to take part in both studies, but only once in each.

Experiment platform

A prototype listening test platform at the Swedish national research infrastructure Språkbanken Tal was used. The platform is fully WCAG 2.1 [24] compliant and presents a single stimuli (sound file) per page. Listeners were guided through their test and then returned to the Prolific web site. Only a very small number of listeners (<3%) timed out or returned their task undone.

Texts and reference stimuli

Two Swedish texts, **TEXT1** and **TEXT2**, each containing two sentences from a campaign concerning covid-19 information, were used, see Table 1. A recording of the texts was already available in a typical speaking rate, and the same voice talent was employed to rerecord the sentences at a very slow speaking rate. The results of these slow recordings were used as references (**REF1** and **REF2**).

Table 1. *Number of sentences, words, syllables and minor delimiters in the two texts.*

Text	Sentences	Words	Syllables	Minor delimiters
TEXT1	2	31	57	1
TEXT2	2	33	66	2

Stimuli

The human stimuli were based on the human recordings reading **TEXT1** and **TEXT2** at a typical speaking rate and at a slow speaking rate. The duration and articulation rates of these files are shown in Table 2. To illustrate the temporal aspects of the texts using the synthetic voice used in the stimuli creation, the data from a synthesised reading with pauses at major and minor delimiters is included in the table.

Table 2. *Duration (seconds) and articulation rate (syllables/second) for the human recordings and synthesis (with pauses at major and minor delimiters).*

	TEXT1		**TEXT2**	
	Dur.	**Art. rate**	**Dur.**	**Art. rate**
Human normal	17	2,08	18,7	1,97
Human slow (**REF**)	30,5	3,90	32,5	3,58
TTS	20,4	3,13	22,7	2,97

Table 3. *A description of the eight stimuli used in the study: pause placements, prolongation (***Stretched** *means the stretching of the whole file,* **Pauses** *the prolongation of non-specch and* ***Both*** *combines* ***Pauses*** *and* **Stretched***), Number of pauses, proportions of non-speech, and average pause durations for both texts.*

	Pause placements	Pro-longation	Number of pauses		Non-speech (%)		Avg. pause duration	
			Text1	Text2	Text1	Text2	Text1	Text2
Ref	HumSlow	None	16	15	19,67	20,62	330	381
HumStretched	Human	Stretched	6	7	11,80	14,15	457	507
TTSStretched	Orthographic	Stretched	2	4	10,16	13,85	925	792
TTSOrthoPauses	Orthographic	Pauses	2	4	35,08	38,46	4933	2895
TTSDesignedPauses	Designed	Pauses	8	7	43,93	40,31	1600	1736
TTSWordPauses	Word	Pauses	30	32	37,05	34,46	342	309
TTSHumPauses	HumSlow	Pauses	16	15	40,00	39,08	710	781
TTSHumPausesStretched	HumSlow	Both	16	15	22,95	21,54	360	382

Ref1 and **Ref2** were of about 30 seconds duration each, and in order to eliminate effects of durational variation in the evaluation, all manipulated stimuli were made to match these durations. **HumStretched1** and **HumStretched2** were created by stretching the typical human recordings to the same duration as the slow readings.

For the synthesized stimuli, we trained a voice with Nvidia's PyTorch implementation of Tacotron and WaveGlow on nearly 18 hours of female speech data from the same voice talent as the recorded data in **Ref1** and **Ref2**, originally recorded for unit selection synthesis [25][26]. The words in the training data were split into five relative speaking rate categories. To ensure there were enough speech data in each category, they were balanced to contain approximately the same number of words (27 000). Each word was prepended with its speaking rate category in the training. This makes it possible to synthesize at five different speaking rates, by inserting the speaking rate category before each word in the input to the synthesizer. The slowest speaking rate, along with hyper-articulated phonemic transcriptions, was used for all synthetizations in this study. Note that the slow synthetic data created in this manner is not nearly as slow as **Ref1** and **Ref2**, simply because the voice is not trained on deliberately slow speech (see Table 2).All stimuli are available at http://www.sprakbanken.speech.kth.se/surveys/slow/.

Four different *pause placements* were used: **Orthographic** pauses were inserted at major and minor delimiters in the orthography (e.g. commas and stops); **Designed** pauses were inserted at selected syntactic boundaries aiming for equally-sized speech chunks (other policies are possible); **Word**, where pauses were inserted between all words; and finally **HumanSlow**, where we copied the locations of perceptual pauses (>120 ms of non-speech [27]) in the **Ref1** and **Ref2**. The initial pause durations were what came out of the synthesis, and all versions were still shorter than the corresponding **Ref1** and **Ref2**.

The stimuli were synthesized with pauses in the locations described, and the duration of each pause was manually manipulated by inserting (or sometimes deleting) silence copied from the same file. For the pause locations in **TTSOrthoPauses**, **TTSDesignedPauses** and **TTSHumPauses** the pause durations were altered to match the proportion of the same pause location in the **Ref** files. For **TTSWordPauses**, we kept the original pause durations between each word from the synthetization, and manipulated the pause locations that also occurred in the **Ref** files proportionally, to end up at the file durations of the **Ref** files. Finally, **TTSHumPausesStretched** were first given the same pause durations as the **Ref** files, then the entire files were stretched to the required durations. The details of the resulting stimuli are presented in Table 3, and a visualization of the speech/non-speech patterns of the readings of **Text1** is shown in Figure 1.

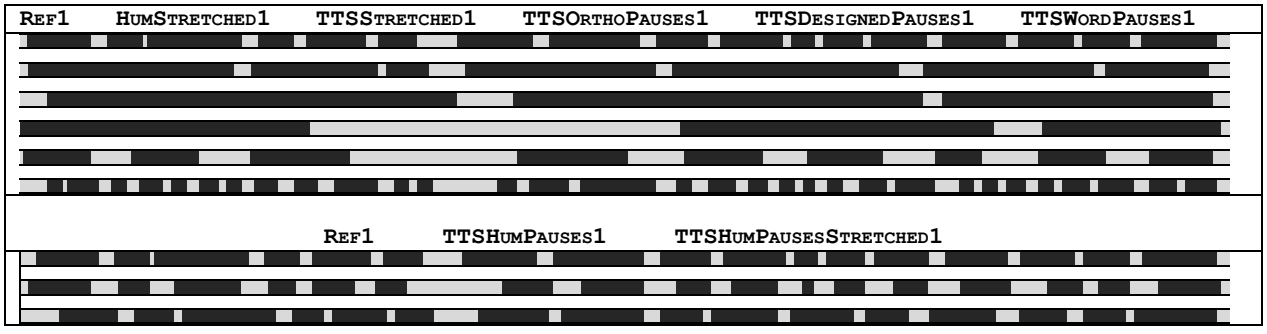

Figure 1. *Chronogram of speech (black segments) and non-speech (grey segments) for Text1.First section from top to bottom:* **Ref1**, **HumStretched1**, **TTSStretched1**, **TTSOrthoPauses1**, **TTSDesignedPauses1**, **TTSWordPauses1**. *Second section:* **Ref1** *(repeated),* **TTSHumPauses**, **TTSHumPausesStretched**.

Procedure

In order to avoid overly long sessions and listening fatigue, we divided the stimuli in two different tests, each containing all 8 versions of one of the two texts. For each test, the order of the stimuli was varied systematically. A single listener could participate in both tests, but only once per test. For each stimulus, they were presented with the framing sentence "Imagine that you have requested a short text to be read for you *very slowly*." and the question "How well does this reading match your expectations?". The response alternatives were a five-grade scale with the option "Matches very well", "Matches well", "Matches neither well nor poorly", "Matches poorly" and "Matches very poorly". We use **ACCEPTANCE** for this variable.

4. Analysis & results

102 recruits started the listening study. 3 did not finish, and the experiment was stopped when 8 recruits had responded to each test set in each systematically varied order. The listening times varied between 5 and 6 minutes. 29 listeners participated in both studies (i.e. judged both texts once) and 70 took part in one study only.

A one-way ANOVA showed statistically-significant difference in **ACCEPTANCE** by stimuli identity ($f(15)=12.088$, $p < 0.001$). Pairwise comparisons showed no significant difference within any pair of the same stimuli type but different texts. This let us combine **TEXT1** and **TEXT2**, so that we considered only stimuli type – the manner in which the stimulus was created. One-way ANOVA again showed statistically-significant difference in **ACCEPTANCE** by stimuli type ($f(7)=22.664$, $p < 0.001$).

Post-hoc pairwise t-tests using the Bonferroni correction were performed across all pairs of stimuli types. There was a significant difference between **REF** and all other stimuli types. In all, 14 pairs were significant. These pairs and the effect sizes are presented in Table 4.

For good measure, the difference between the two human readings (**REF** and **HUMSTRECTCHED**) and the other stimuli were verified with Dunnett's test for comparing several treatments with a control. All synthesized stimuli were significantly different than **REF** at the 0.001 level, and only **TTSORTHOPAUSES** was differed from **HUMSTRETCHED**, again at the .001 level.

Finally, we performed a Tukey's test for all pairs of stimuli type. This singled out the same 14 pairs as significantly different, at the same levels as the repeated t-tests with Bonferroni correction.

The literature, our intuition from listening to the stimuli, and the initial results all hint at a combination of the duration of talkspurts and their frequency as being key to slow speech (note that the speech/pause ratio and other similar metrics can be derived from these two measures). As talkspurts can clearly be both too long and too short, and their frequency too high or too low, quadratic polynomials we fitted to their averages (**AVTSDUR** and **AVTSFREQ**) for all stimuli. As expected, both significantly predict **ACCEPTANCE**, and there are interaction effects. The additive model's F statistic is 18.181*** (df = 4; 1019), and the corresponding multiplicative model yields 14.561*** (df = 8; 1015). Adjusted R2 is 0.063 and 0.096, respectively.

Table 4. *Each row describes the stimuli listed in the leftmost column, starting with the number of judgements, the average, and the standard deviation. The last three columns contain the significant pairwise comparisons, with (1) representing* **REF**, *(2) representing* **TTSORTHOPAUSES**, *and (3)* **TTSDESIGNEDPAUSES**. *Each cell shows effect size and significance (0,05=*, 0.01=**, 0.005=***,0.001=****)*

	N	Avg	SD			
REF (1)	128	4,2	1,0	**(1)**	**(2)**	**(3)**
HUMSTRETCHED	128	3,1	1,4	-0.9 (L) ****	0.50 (M) **	-
TTSSTRETCHED	128	3,2	1,3	-0.91 (L) ****	0.61 (M) ****	-
TTSORTHOPAUSES (2)	128	2,4	1,2	-1.6 (L) ****	n/a	-
TTSDESIGNEDPAUSES (3)	128	2,8	1,2	-1.3 (L) ****	-	n/a
TTSWORDPAUSES	128	3,3	1,2	-0.85 (L) ****	0.69 (M) ****	0.41 (S) *
TTSHUMPAUSES	128	3,2	1,1	-0.96 (L) ****	0.66 (M) ****	-
TTSHUMPAUSES STRETCHED	128	3,4	1,2	-0.75 (M) ****	0.80 (M) ****	0.51 (M) **

5. Discussion

The reference readings score higher than all other readings on the question of how well it corresponds to expectations of a very slow reading. This is to be expected, not only because it is read by a human professional who has been instructed to read very slowly, but because none of the other readings are designed, originally, to create very slow speech.

Compared to the typically-paced and stretched human reading **HUMSTRETCHED**, only two of the TTS varieties perform significantly worse: **TTSORTHOPAUSES**, in which only the very few orthographic pauses (commas and full stops) are extended to reach the duration of the reference utterances. These readings were included in part as a test case to see that the crowd workers behaved as could be expected, and in part to highlight the fact that pause lengthening has an upper bound. **TTSORTHOPAUSES** is judged as significantly worse than 6 out of the 7 other readings. Finally the reading with pauses inserted between constituents at regular intervals, **TTSDESIGNEDPAUSES**, fares poorly against the two highest ranked TTS readings, but the effect is small.

Turning to the average scores, the reference utterances stand out with a 4.2 average, as does **TTSORTHOPAUSES** with 2.4. The rest of the readings receive scores slightly above 3, with the stretched human voice ending up somewhere in the middle. Having listened to the stimuli, we propose that with the exception of **TTSORTHOPAUSES**, the stimuli are designed to be as pleasant to listen to as possible. The one other exception, perhaps, is **TTSWORDPAUSES**, with a pause between every single word. We did not expect this to be a viable solution, but having listened to the result ourselves, it really does not sound bad.

6. Conclusions & future work

The goal of this study has been to see if acceptable slow speech can be created with relatively simple means, without rerecording databases. Out of six different methods of prolonging synthesised utterances to match, in duration, *very* slow human speech, five achieved the same rating as original human speech that had been stretched. This is promising.

The results suggest that the relation between speech and non-speech play a role: about 10% of the variation in **ACCEPTANCE** is explained by a regression model based on these factors, in spite of the materials being highly varied in nature and not at all varied systematically in terms of speech/non-speech relation. As mentioned in the discussion, the literature supports this finding, and the very poor acceptance of **TTSORTHOPAUSES** is perhaps related to the uncomfortable pauses Sacks et al call "lapse"[28], the minimum duration of which Jefferson and others have approximated to 1 second [29].

We believe that we now have the tools to create workable very slow speech using only moderately slow speech synthesis, by manipulating the placement and durations of pauses, and the next step is a structured study of the relation between talkspurt durations and pause durations.

7. Acknowledgements

The survey was partly funded by Vinnova (2018-02427). The results will be made more widely accessible through the Swedish Research Council funded national infrastructure Nationella språkbanken and Swe-Clarin (2017-00626).

References

[1] 'Talking books', *Swedish Agency for Accessible Media*. https://www.mtm.se/english/products-and-services/talking-books/ (accessed Apr. 28, 2021).

[2] S. E. Blumstein, B. Katz, H. Goodglass, R. Shrier, and B. Dworetsky, 'The Effects of Slowed Speech on Auditory Comprehension in Aphasia', *Brain and Language*, vol. 24, pp. 246–265, 1985.

[3] K. Hux, J. A. Brown, S. Wallace, K. Knollman-Porter, A. Saylor, and E. Lapp, 'Effect of Text-to-Speech Rate on Reading Comprehension by Adults With Aphasia', *Am J Speech Lang Pathol*, vol. 29, no. 1, pp. 168–184, Feb. 2020, doi: 10.1044/2019_AJSLP-19-00047.

[4] J. E. Sung *et al.*, 'Real-time Processing in Reading Sentence Comprehension for Normal Adult Individuals and Persons with Aphasia', *Aphasiology*, vol. 25, no. 1, pp. 57–70, 2011, doi: 10.1080/02687031003714434.

[5] F. Goldman-Eisler, 'The Determinants of the Rate of Speech Output and their Mutual Relations', *Journal of Psychosomatic Research*, vol. 1, pp. 137–143, 1956.

[6] F. Goldman-Eisler, 'The Significance of Changes in the Rate of Articulation', *Lang Speech*, vol. 4, no. 3, pp. 171–174, Jul. 1961, doi: 10.1177/002383096100400305.

[7] E. Jacewicz, R. A. Fox, C. O'Neill, and J. Salmons, 'Articulation rate across dialect, age, and gender', *Lang Var Change*, vol. 21, no. 2, pp. 233–256, Jul. 2009, doi: 10.1017/S0954394509990093.

[8] B. Picart, T. Drugman, and T. Dutoit, 'Analysis and Synthesis of Hypo and Hyperarticulated Speech', in *Speech Synthesis Workshop (SSW)*, Kyoto, Japan, 2010, vol. 28, pp. 687–707, doi: https://doi.org/10.1016/j.csl.2013.04.008.

[9] S. Tauroza and D. Allison, 'Speech Rates in British English', *Applied Linguistics*, vol. 11, no. 1, pp. 90–105, 1990, doi: 10.1093/applin/11.1.90.

[10] T. H. Crystal and A. S. House, 'Articulation rate and the duration of syllables and stress groups in connected speech', *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 101–112, 1990.

[11] J. Trouvain, J. Koreman, A. Erriquez, and B. Braun, 'Articulation Rate Measures and Their Relation to Phone Classification in Spontaneous and Read German Speech', in *Adaptation-2001*, 2001, pp. 155–158.

[12] J. Koreman, 'Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech', *The Journal of the Acoustical Society of America*, vol. 119, pp. 582–96, Feb. 2006, doi: 10.1121/1.2133436.

[13] A. Leemann, 'Analyzing geospatial variation in articulation rate using crowdsourced speech data', *Journal of Linguistic Geography*, vol. 4, no. 2, pp. 76–96, 2016, doi: https://doi.org/10.1017/jlg.2016.11.

[14] G. Fant, A. Kruckenberg, and J. Barbosa, 'Individual variations in pausing. A study of read speech', presented at the Fonetik, Umeå, Sweden, 2003.

[15] J. Dankovicova, 'Articulation Rate Variation within the Intonation Phrase in Czech and English', in *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco, 1999, p. 4.

[16] B. Lindblom, 'Some Temporal Regularities of Spoken Swedish', in *Auditory Analysis and Perception of Speech*, Elsevier, 1975, pp. 387–396.

[17] E. Cooper, E. Li, and J. Hirschberg, 'Characteristics of Text-to-Speech and Other Corpora', in *9th International Conference on Speech Prosody 2018*, 2018, pp. 690–694, doi: 10.21437/SpeechProsody.2018-140.

[18] K. J. Kohler, 'Parameters of Speech Rate Perception in German Words and Sentences: Duration, Fo Movement, and Fo Level', *Lang Speech*, vol. 29, no. 2, pp. 115–139, Apr. 1986, doi: 10.1177/002383098602900202.

[19] A. C. M. Rietveld and C. Gussenhoven, 'Perceived speech rate and intonation', *Journal of Phonetics*, vol. 15, no. 3, pp. 273–285, Jul. 1987, doi: 10.1016/S0095-4470(19)30571-6.

[20] G. Fant, A. Kruckenberg, and L. Nord, 'Temporal organization and rhythm in Swedish', in *Proceedings of the XIIth ICPhS*, Aix-en-Provence, France, 1991, pp. 251–256.

[21] S. Shechtman and A. Sorin, 'Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities', presented at the Speech Synthesis Workshop (SSW 10), Vienna, Austria, 2019, doi: 10.21437/SSW.2019-49.

[22] T. Raitio, R. Rasipuram, and D. Castellani, 'Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features', presented at the Interspeech, Sep. 2020, doi: DOI: 10.21437/Interspeech.2020-2861.

[23] S. Palan and C. Schitter, 'Prolific.ac - A subject pool for online experiments', *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2017.

[24] 'Web Content Accessibility Guidelines (WCAG) 2.1', *W3C*. https://www.w3.org/TR/WCAG21/ (accessed Apr. 28, 2021).

[25] C. Tånnander, 'Speech Synthesis and evaluation at MTM', in *Proceedings of Fonetik*, Gothenburg, Sweden, 2018, pp. 75–80,

[26] C. Tånnander and J. Edlund, 'Stress manipulation in text-to-speech synthesis using speaking rate categories. Fonetik, Lund, Sweden, submitted.

[27] M. Heldner, 'Detection thresholds for gaps, overlaps, and no-gap-no-overlaps', *JASA*, vol. 130, no. 1, pp. 508–513, Jul. 2011, doi: 10.1121/1.3598457.

[28] 'A Simplest Systematics for the Organization of Turn Taking for Conversation', in *Studies in the Organization of Conversational Interaction*, Academic Press, 1978, pp. 7–55.

[29] G. Jefferson, 'Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation', in *Conversation: An interdisciplinary perspective*, Clevedon, England: Multilingual Matters, 1989, pp. 166–196.