ESUKA – JEFUL 2019, 10–1: 167–190

TOWARDS A VIVIDNESS IN SYNTHESIZED SPEECH FOR AUDIOBOOKS

Hille Pajupuu[1], Rene Altrov[1], and Jaan Pajupuu[2]
[1]Institute of the Estonian Language and [2]Industry62

Abstract. The goal of this study was to determine which acoustic parameters are significant in differentiating the speaking styles of a narrator and that of male and female characters as voiced by a reader of audiobooks. The study was initiated by a need to improve the expressivity and differentiation of speaking styles in fiction books read out by synthesized voices. The corpus used as research material was created from an audio novel, as read by a professional male voice artist. To determine whether it is possible to identify these speaking styles from the voice of the reader, a web-based perception test consisting of 48 sentences was conducted. The results showed that the listeners identified all three styles. For acoustic analysis, the openSMILE toolkit was used and 88 eGeMAPS-defined parameters were extracted for every sentence in the corpus. All styles were differentiated by 38 statistically significant parameters. To improve vividness, synthesizers aimed at reading fiction books could be trained to perform all three styles.

Keywords: audiobooks, speaking style, direct speech, characters' speech, GeMAPS, speech analysis, expressive speech synthesis

DOI: https://doi.org/10.12697/jeful.2019.10.1.09

1. Introduction

Audiobooks offer good source material for creating corpora for speech synthesis. They include a single speaker talking in a variety of styles and with varying expressivity while offering a high-quality audio signal with corresponding text. The creation of audiobook corpora has prioritized different aspects, depending on the method of speech synthesis and the text type for which the synthesizer is intended. For a synthetic voice suitable for reading different types of text (e.g., news reports, reviews, personal blogs), it may be necessary to leave out direct speech, as its voice parameter variability is too high, whereas for the reading of one single type of expressive text (e.g., fiction), it may be

beneficial to label the speech in audiobooks by the different levels of expressivity of speaking styles and to train the synthesizer to produce them all (see, e.g., Chalamandaris et al. 2014, Charfuelan and Steiner 2013, Chistikov et al. 2014, Eyben et al. 2012, and Sini et al. 2018).

If the goal of speech synthesis is to improve the vividness of the performance of fiction audiobooks, then the key issue is how to label audiobook corpora so that the resulting synthesis is similar to a human performance and appropriate to the listeners' cultural expectations. Professional audiobook voice artists can use different speaking styles depending on genre, personal preference, and the demands placed on them: they may choose to personify every character using their voice (especially for children's books) or only differentiate narrative parts from dialogue (see, e.g., Alain et al. 2017, Mihkla et al. 2018, Montaño et al. 2013, and Zhao et al. 2006). Montaño and Alías (2017) have shown that the content of a book determines the mode of reading aloud, regardless of language. Yet it is clear that there exist culturally different expectations for the paralinguistic aspects of an audiobook (see, e.g., Stolarski 2017a, 2017b).

Until now, a purpose-built, carefully formed neutral speech corpus has been used for Estonian language speech synthesis (Piits 2016). Some research has been done to find out how synthesizers based on this corpus could emphasize direct speech in texts.

In perception tests carried out by Mihkla et al. (2017), it was found that listeners of audiobooks read by synthesizers preferred direct speech to be produced at 2.5 semitones higher and 3 dB louder than reporting clauses.

Mihkla et al. (2018) investigated how professional audiobook readers change their pitch to differentiate narrative passages, direct speech, and reporting clauses. Based on four passages of 12–29 minutes, they ascertained, using a fundamental frequency (F0) baseline, that direct speech and reporting clause F0 baselines only differed significantly if the reporting clause followed direct speech or if the reporting clause was between continuing direct speech. Reporting clauses were characterized by a lower F0 baseline compared to direct speech. In Spanish, a lower pitch (F0 mean) and intensity have also been noticed in post-character sentences (Montaño et al. 2013, Montaño and Alías 2016). A comparison of Spanish, German, French, and English has shown that

post-character sentences show similar acoustic distributions across all languages: they tend to be spoken with a muffled voice that implies lower F0 and intensity values (Montaño and Alías 2017).

Research by Mihkla et al. (2013, 2014) dealt with the voicing of subtitles of films shown on television. There is a necessity to change the parameters of a specific synthesized voice when a single frame shows subtitles containing the speech of multiple people and they must be differentiated. To achieve this, they used changes in F0, tempo, and timbre. The speaking turn of a new speaker is marked in subtitles, but not the gender of the speaker (the viewer being aided by the video), necessitating that the voice suit both female and male characters. The researchers concluded that only a slight change of parameters is sufficient (not exceeding 10% of the initial parameters of the voice).

Despite interlocutors being easier to identify in a book of fiction (due to occasional reporting clauses) compared to subtitles, text-based automatic character identification remains a non-trivial problem (e.g., Zhang et al. 2003, Elson and McKeown 2010, He et al. 2013, and Iosif and Mishra 2014). Currently a synthesizer cannot always convincingly voice a fiction book without audiobook editors who help note character turns and define male and female speech.

Stemming from a need to improve the vividness and differentiation of narrative, female, and male character speech, we set out to identify the acoustic means used, their significant features, and their effectiveness on the basis of audiobooks. For this we analysed an audio novel recorded by a professional male voice artist. Listening to the novel, we noted three different speaking styles: the narrator's speech (uses narrative and descriptive mode[1], does not include dialogues) and male and female character direct speech (dialogues). Of all characters, the voice artist identifiably characterized only one male and one female character.

To achieve our aims, we established the following research questions:

1. Do listeners distinguish among the different speaking styles of the voice artist's reading of an audio novel: the narrator's speech and the male and female characters' direct speech?

1 The narrative mode is generally used to inform the listener/reader about the actions that are taking place in the story, whereas the descriptive mode has the function of describing characters, environments, objects, etc. (Montaño et al. 2016).

2. What are the acoustic features distinguishing narrator's speech, male character direct speech, and female character direct speech in an audio novel?

As far as we know, earlier studies have not ascertained the differentiating parameters for these three speaking styles.

When asked to imitate a feminine and masculine voice, people imagine that female voices sound higher and masculine voices lower. This was ascertained by a study by Cartei et al. (2012), where 31 native British-English adult speakers were asked to read first using their normal voice and then while sounding as masculine and feminine as possible. Both men and women raised the pitch and formant frequencies (mainly responsible for the timbre) of their voices when feminizing their voice and lowered them when masculinizing their voice.

Studies by Stolarski (2017a, 2017b) on the reading out of literary characters' direct speech have considered fundamental frequency along with intensity and their variability (standard deviation). Results showed that the speech of female characters was either read out with a slightly higher (4.7%) F0 in comparison to the participants' own mean F0 or there was no change. Only two male readers and two female readers out of 64 raised their fundamental frequency by a significant amount (35.9% to 58.9%). For the speech of male characters, fundamental frequency was not lowered. There was no difference between female and male characters in fundamental frequency variability. Unlike American-English readers, British-English readers had a tendency to raise the fundamental frequency when reading dialogues, regardless of the characters' gender. Studying intensity and its variability, Stolarski (2017a, 2017b) found that the characters' gender had no effect. Yet it was found that male American-English readers tended to lower their intensity when reading dialogues.

Studies on direct and indirect speech in reading out fiction books have shown that the tempo of direct speech varies more than indirect speech (Yao and Scheepers 2011, 2015). The tempo of direct speech depended on context, whereas for indirect speech context had no effect on reading tempo. Their study also showed that fundamental frequency varied more in direct speech, making it more vivid, whereas indirect speech was more neutral and less varied.

In our study we investigated the interactions among many of the acoustic parameters of narrator's speech and female and male characters' speech, to determine the style of performance characteristic of Estonian culture.

2. Method

2.1. Material

For analysis we used a corpus created from the audio novel *Tõde ja õigus I* [Truth and Justice I] by Anton Hansen Tammsaare (size 10,223 sentences, 21.1 hours), read out by a male professional voice artist. We labelled the narrator's speech (5,619 sentences), male and female characters' direct speech (1,516 and 519 sentences, respectively), and male and female characters' direct speech with reporting clauses. The latter group was left out of this study.

2.2. Perception test

To ascertain whether a listener can distinguish the different speaking styles in the voice of the male voice artist, we arranged a web-based perception test. A group of 12 men (aged 24–64) and 11 women (aged 20–65) listened to 48 sentences chosen from the audiobook corpus. Of these, 16 were the narrator's speech sentences, 16 were female character speech sentences, and 16 were male character speech sentences. The sentences were chosen to not include clues about the speaker, being appropriate for the speech of the narrator or for female or male characters (e.g., *Mis inimene külvab, seda tema ka lõikab. [You reap what you sow.]*; *Et kui nad kahekesi seal hästi läbigi saaks, aga seda ka ei ole. [If the two at least got along there, but even that is not given.]*) The listeners were asked whose speech they had heard, with three options: the narrator, a female character, or a male character.

2.3. Acoustic analysis

For acoustic analysis, we used all the narrator's, male characters', and female characters' speech sentences in the corpus. The analysis was carried out using the openSMILE v2.3.0 toolkit (Eyben et al. 2013).

A total of 88 parameters were extracted for each sentence, which are defined by eGeMAPS (the extended Geneva Minimalistic Acoustic Parameter Set) as the standardized set of acoustic speech parameters for various areas of automatic voice analysis, including paralinguistic speech analysis. eGeMAPS includes parameters for speech frequency, energy and amplitude, spectral characteristics (balance), and tempo (Eyben et al. 2016).

In addition, using Lindh and Eriksson's (2007) formula, the voice artist's F0 baseline when reading narrator, female, and male character speech was calculated:

$$F0b = F0mean - k \cdot sd(F0),$$

where
F0b = baseline;
F0mean = mean fundamental frequency (Hz);
K = empirically derived constant 1.43; and
sd(F0) = standard deviation of fundamental frequency.

To find the acoustic parameters differentiating narrator and male and female character sentence groups, the R program was used (R Core Team 2017). To detect the effect of the group on the parameters, the raw values for each parameter were scaled over all groups using the *scale(data, centre = TRUE, scale = TRUE)* method.

We used non-parametric methods because some of the parameters were not normally distributed. The Kruskal–Wallis test "kruskal.test" was used to discover the parameters with significantly different group medians. For these parameters, we further applied the "pairwise.wilcox.test" to find out which groups were significantly different.

Confidence intervals (95%) were calculated for the median values of all three groups for each parameter. Then we used the group median CI range to select parameters for classifying the sentences. If the group CI range was fully above zero or fully below zero, the parameter was considered distinctive for this group.

3. Results

3.1. Perception test

The listeners identified the narrator's speech and female and male characters' direct speech from the voice of the male voice artist reading an audiobook. The answers given (total 1,061) are presented as percentages in Table 1.

Table 1. Confusion matrix for identifying speaking style groups: narrator's speech, male and female characters' direct speech.

Listener answer / Speaking style	Narrator	Male character	Female character
Narrator	**58**	30	12
Male character	15	**60**	25
Female character	14	28	**58**

Note. The numbers are the choice percentages for every speaking style. The diagonal shows correctly identified speaking styles.

Table 1 reveals that participants identified all three speaking styles. None of the incorrect answers exceeded chance probability (33.3%). Listeners were more likely to confuse male and female characters' direct speech with one another, rather than direct speech with narrator's speech.

3.2. Acoustic analysis

Out of the 88 eGeMAPS parameters, 38 parameters statistically significantly differentiated all three speaking styles: eight frequency-related parameters, 11 energy-/amplitude-related parameters, 18 spectral (balance) parameters, and one tempo parameter (see Table 2).

Table 2. Parameters differentiating all speaking styles.

eGeMAPS parameter and description	Parameter group	Kruskal-Wallis test statistic	↓	=	↑
F0semitoneFrom27.5Hz_sma3nz_percentile50.0 50th percentile of fundamental frequency (F0) on a semitone frequency scale	FRQ	536.8****	N	M	F
spectralFluxV_sma3nz_amean Mean of spectral flux (difference of the spectra of two consecutive frames) of voiced regions	S	513.5****	N		F, N
F0semitoneFrom27.5Hz_sma3nz_amean Mean of F0 on a semitone frequency scale	FRQ	474.6****	N		M, F
F0semitoneFrom27.5Hz_sma3nz_percentile80.0 80th percentile of F0 on a semitone frequency scale	FRQ	465.9****	N		M, F
hammarbergIndexV_sma3nz_amean Mean of Hammarberg index (ratio of the strongest energy peaks in the 0–2 kHz vs 2–5 kHz regions) of voiced regions	S	448.8****	M	F	N
F3frequency_sma3nz_amean Mean of the third formant (F3) frequency	FRQ	444.2****	N	F	M
F0semitoneFrom27.5Hz_sma3nz_percentile20.0 20th percentile of F0 on a semitone frequency scale	FRQ	440.4****	N		M, F
F2frequency_sma3nz_amean Mean of the second formant (F2) frequency	FRQ	381.0****	N, F		M
loudness_sma3_pctlrange0.2 Range of the 20th to 80th percentile of loudness	E/A	367.6****	N	F	M

eGeMAPS parameter and description	Parameter group	Kruskal-Wallis test statistic	↓	=	↑
logRelF0.H1.A3_sma3nz_amean Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)	E/A	366.1****	M	F	N
alphaRatioV_sma3nz_amean Ratio of energy above 1 kHz (up to 5 kHz) to energy below 1 kHz, voiced segments	S	360.3****	N		F, M
loudness_sma3_stddevRisingSlope Standard deviation of the rising slopes of loudness	E/A	340.1****	N	F	M
loudness_sma3_percentile80.0 80th percentile of loudness	E/A	331.4****	N	F	M
mfcc2V_sma3nz_amean Mean of mel-frequency cepstral coefficient 2 of voiced regions	S	320.6****	M		N, F
slopeV0.500_sma3nz_amean Mean of spectral slope 0–500 Hz (linear regression slope of the logarithmic power spectrum)	S	297.0****	N		M, F
HNRdBACF_sma3nz_amean Mean of harmonics-to-noise ratio	E/A	240.5****	M, N		F
StddevVoicedSegmentLengthSec Standard deviation of voiced segment length in seconds	T	240.3****	F, M, N		
mfcc4V_sma3nz_amean Mean of mel-frequency cepstral coefficient 4 of voiced regions	S	235.5****	M	F	N
loudness_sma3_meanFallingSlope Mean of the falling slopes of loudness	E/A	221.1****	N	F	M
equivalentSoundLevel_dBp Sound level (RMS converted to decibel with 10*log10(x))	E/A	217.7****		N	F, M

eGeMAPS parameter and description	Parameter group	Kruskal-Wallis test statistic	↓	=	↑
mfcc1V_sma3nz_amean Mean of mel-frequency cepstral coefficient 1 of voiced regions	S	203.0****	F, M		N
mfcc3_sma3_amean Mean of mel-frequency cepstral coefficient 3	S	186.4****	F, M		N
F0semitoneFrom27.5Hz_sma3nz_pctlrange0.2 Range of the 20th to 80th percentile of F0 on a semitone frequency scale	FRQ	166.8****	N	M	F
shimmerLocaldB_sma3nz_amean Mean difference of the peak amplitudes of consecutive F0 periods	E/A	163.1****	F, M		N
mfcc3_sma3_stddevNorm Standard deviation of mel-frequency cepstral coefficient 3	S	162.4****	N, M		F
spectralFlux_sma3_amean Mean of spectral flux (difference of the spectra of two consecutive frames)	S	157.1****	N	F	M
F1amplitudeLogRelF0_sma3nz_stddevNorm Standard deviation of F1 to F0 relative energy	E/A	152.9****	N		M, F
slopeUV500.1500_sma3nz_amean Mean of spectral slope 500–1500 Hz (linear regression slope of the logarithmic power spectrum)	S	142.4****	M, F		N
loudness_sma3_amean Mean of loudness	E/A	135.8****	N	F	M
mfcc3V_sma3nz_amean Mean of mel-frequency cepstral coefficient 3 of voiced regions	S	134.8****	F	M	N
mfcc3V_sma3nz_stddevNorm Standard deviation of mel-frequency cepstral coefficient 3 of voiced regions	S	122.3****	N, M, F		

eGeMAPS parameter and description	Parameter group	Kruskal-Wallis test statistic	↓	=	↑
mfcc1_sma3_amean Mean of mel-frequency cepstral coefficient 1	S	115.2****	M, F		N
mfcc2_sma3_amean Mean of mel-frequency cepstral coefficient 2	S	101.9****	M		N, F
mfcc2_sma3_stddevNorm Standard deviation of mel-frequency cepstral coefficient 2	S	91.5****	N, M, F		
alphaRatioUV_sma3nz_amean Mean of ratio of the summed energy from 50–1000 Hz and 1–5 kHz of unvoiced regions	S	74.3****	M	F	N
mfcc2V_sma3nz_stddevNorm Standard deviation of mel-frequency cepstral coefficient 2 of voiced regions	S	71.4****	N, M, F		
F2amplitudeLogRelF0_sma3nz_stddevNorm Standard deviation of F2 to F0 relative energy	E/A	53.2****	N	M	F
jitterLocal_sma3nz_amean Mean difference in individual consecutive F0 period lengths	FRQ	17.1***	N, F	M	

Note. Parameter groups: FRQ = frequency-related parameters, E/A = energy-/amplitude-related parameters, S = spectral (balance) parameters, T = tempo parameters. Suffixes *_sma3* and *_sma3nz* mean that parameter is smoothed over time with a symmetric moving average filter 3 values long; suffix *stddevNorm* means that stddev is normalized by dividing it by the value of arithmetic mean; suffix *nz* means non-zero values only (Eyben et al. 2016); High (↑), medium (=), and low (↓) denote speaking style groups, which have parameters with CI range of median fully above 0, includes 0, or fully below 0, respectively. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; if in a single column, then sorted in increasing order of median absolute value. *** $p < .001$, **** $p < .0001$.

Figures 1–6 present those parameters which are most used in studies that have analysed narrator's and/or characters' direct speech or studied how readers imitate female and male voices. Among eGeMAPS parameters, these are the mean F0 (characterizes pitch); standard deviation of F0 (characterizes pitch variability); mean loudness; variability in loudness; voiced segments per second (characterizes speech tempo); and the standard deviation of voiced segment lengths (characterizes variability of speech tempo).

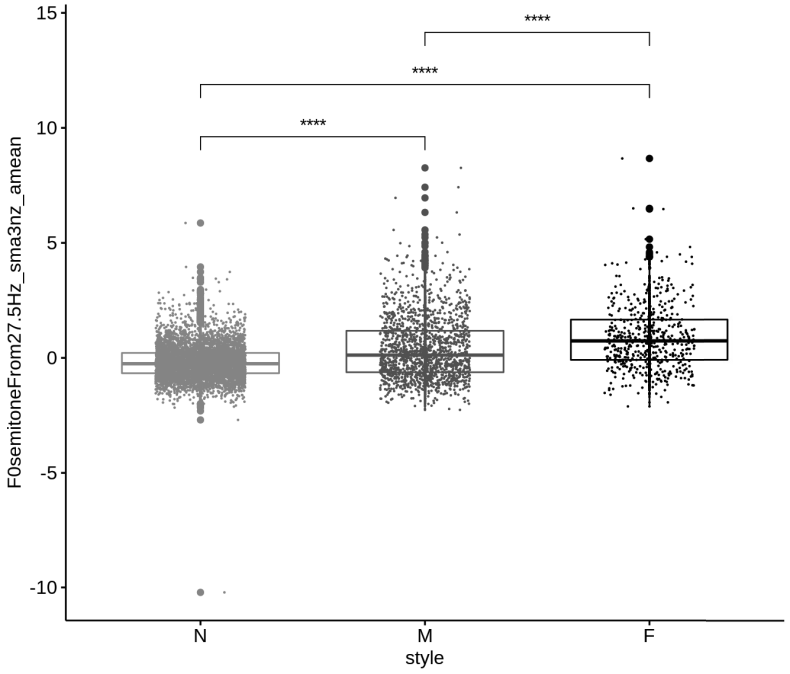

Figure 1. Pitch: Mean of F0. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; **** $p < .0001$.

Figure 1 shows that the three speaking styles differ significantly by pitch. The narrator's speech is the lowest and the female characters' speech is the highest (see also Table 2).

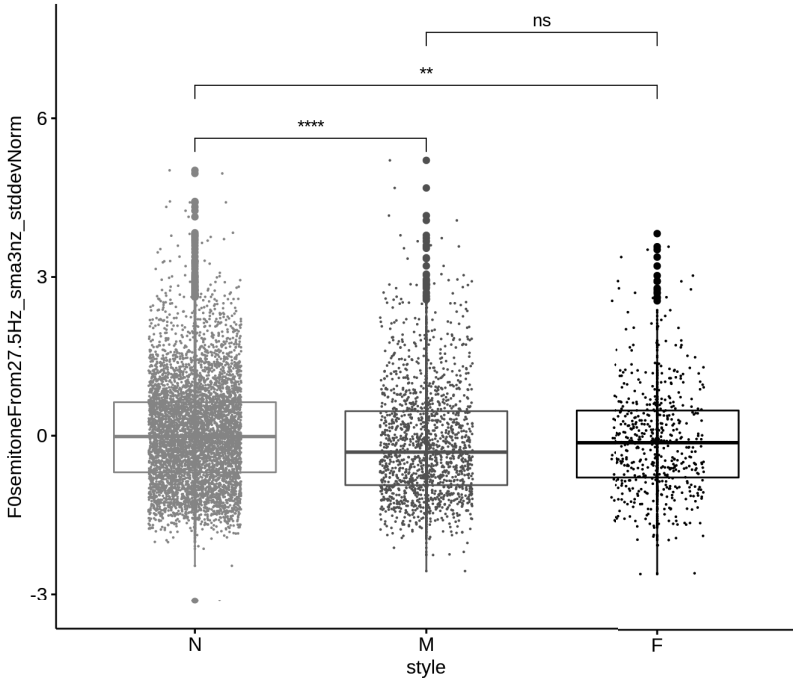

Figure 2. Pitch variability: SD of fundamental frequency. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; ** $p < .01$, **** $p < .0001$, ns = nonsignificant ($p > .05$).

Figure 2 shows that male and female characters' speech did not differ significantly in pitch variability. Yet the narrator's speech pitch variability differed from both gendered characters. The pitch variability was higher for the narrator's speech as read by this male voice artist.

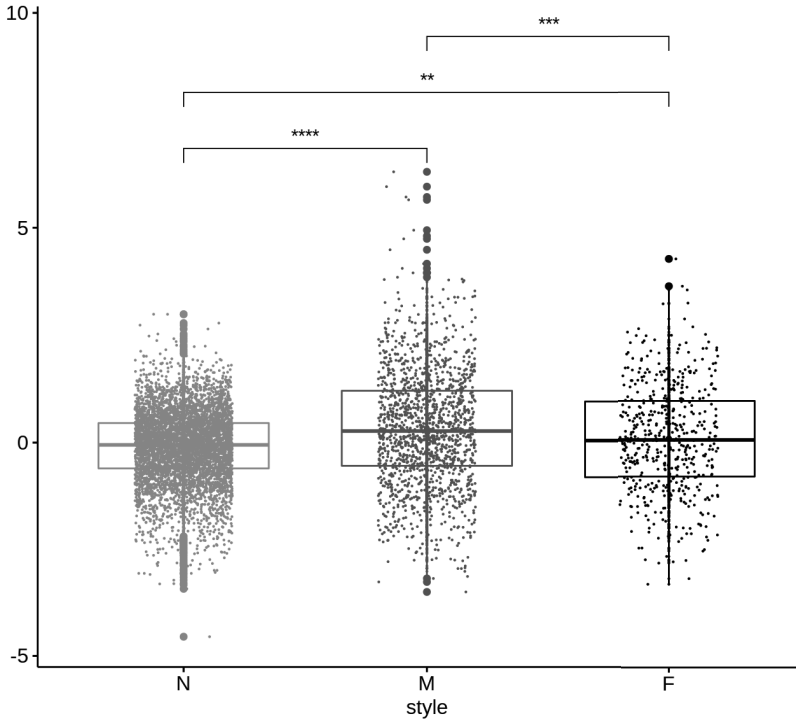

Figure 3. Mean of loudness. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; ** $p < .01$, *** $p < .001$, **** $p < .0001$.

Figure 3 shows that the mean loudness of all speaking styles varied significantly from one another. The narrator's speech was the most quiet, and the male characters' speech was the loudest (see also Table 2).

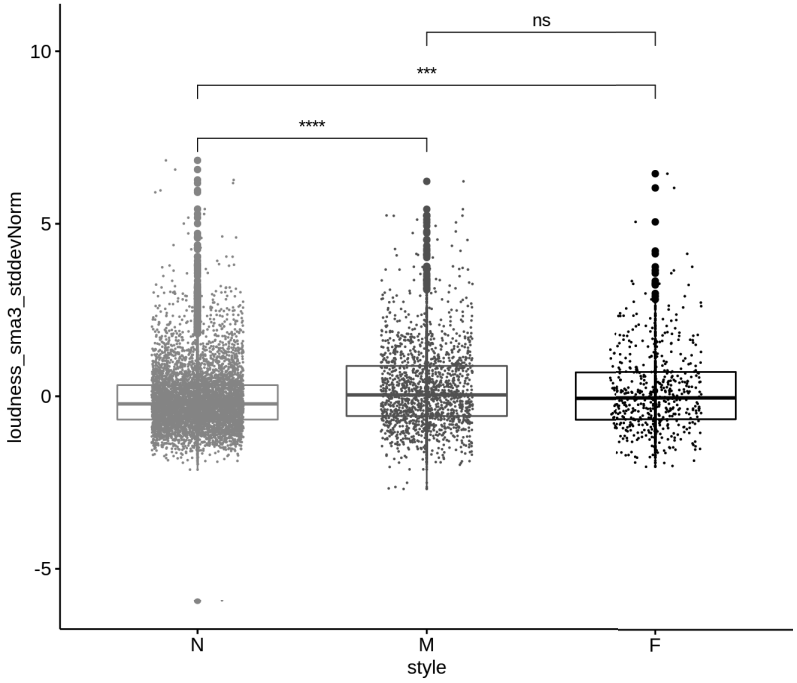

Figure 4. Variability in voice loudness: SD of loudness. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; *** $p < .001$, **** $p < .0001$; ns = nonsignificant ($p > .05$).

Figure 4 shows that variability in voice loudness for male and female characters' speech did not differ significantly. Yet there were differences between the loudness variability of both from the loudness variability of the narrator's speech. The narrator's speech was less varied in loudness compared to direct speech.

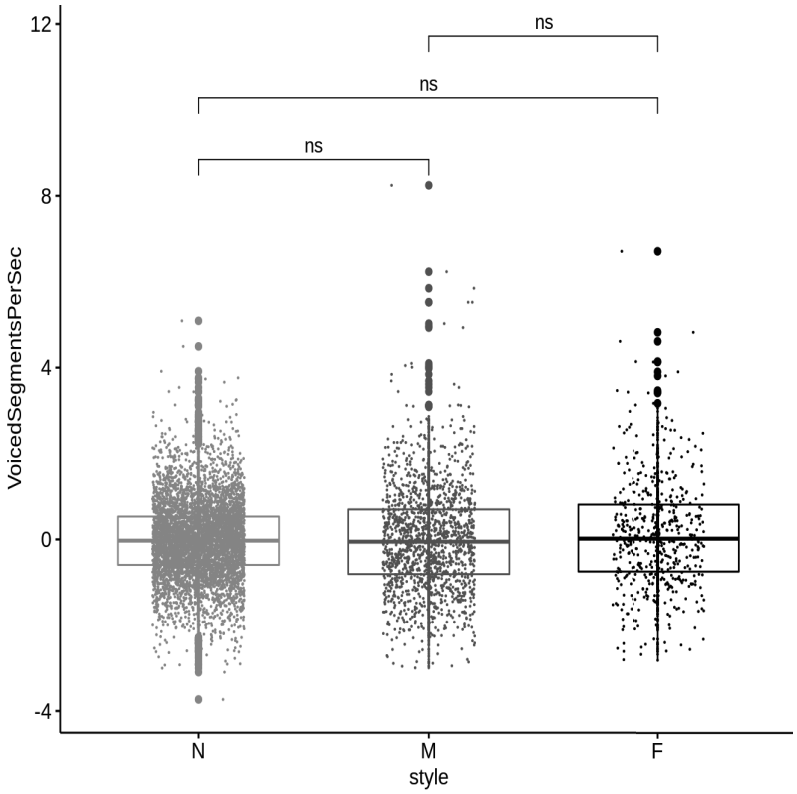

Figure 5. Speech tempo: voiced segments per second. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; ns = nonsignificant ($p > .05$).

Based on Figure 5, we can claim that different speaking styles did not differ significantly in tempo.

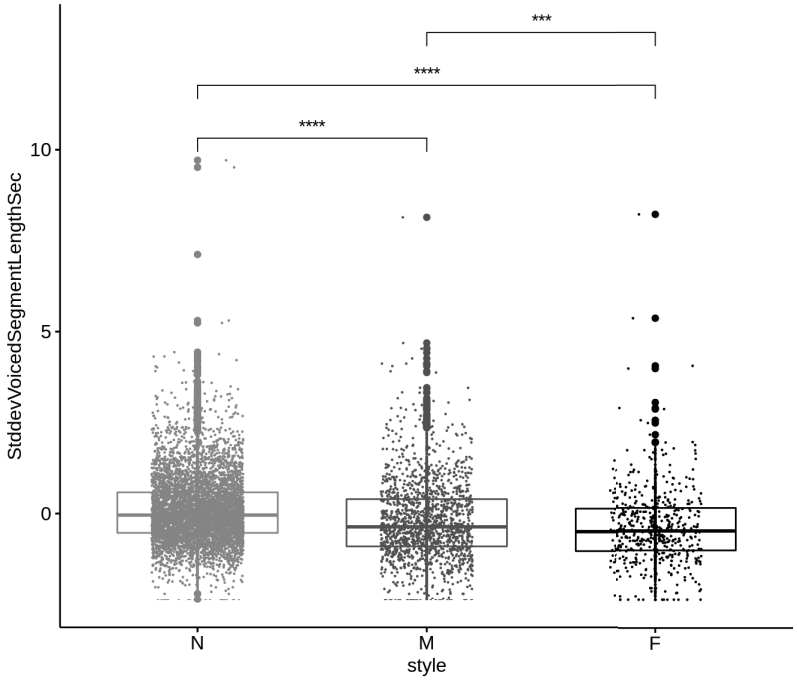

Figure 6. Variability in speech tempo: SD of voiced segments' length in seconds. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; *** $p < .001$, **** $p < .0001$.

Figure 6 shows that the variability of speech tempo significantly differentiated all speaking styles. Female characters' speech tempo varied the least, and the narrator's speech tempo varied the most (see also Table 2).

In addition to eGeMAPS parameters, an F0 baseline was calculated for every speaking style (see Table 3 and Figure 6).

Table 3. Speaking styles F0 baseline in Hz.

Character	1st Qu	Median	3rd Qu
Narrator	62.1	69.2	74.3
Male character	65.5	72.2	78.7
Female character	65.3	73.5	81.6

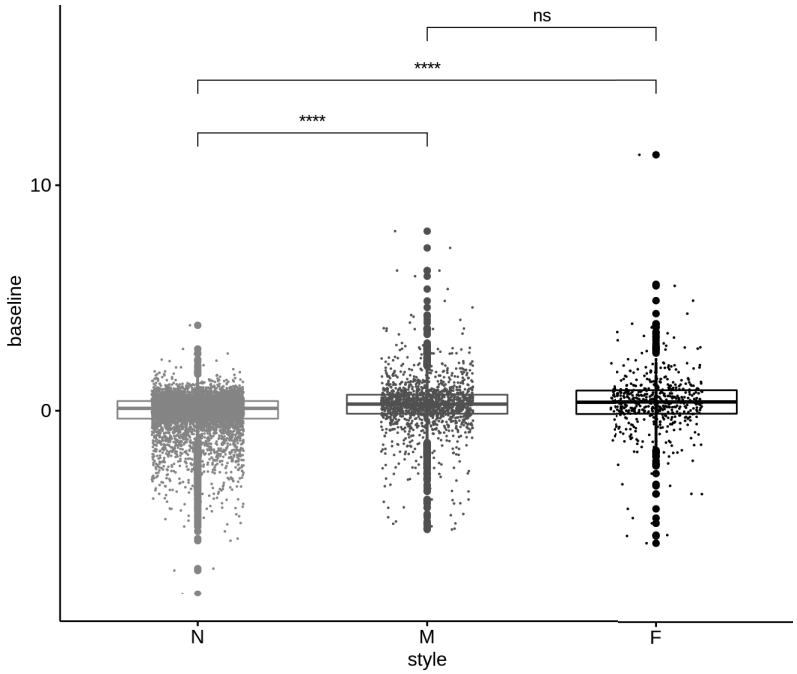

Figure 7. F0 baseline. Speaking styles: N = narrator's speech, M = male characters' speech, F = female characters' speech; **** $p < .0001$; ns = nonsignificant ($p > .05$).

Figure 7 and Table 3 show that the narrator's speech F0 baseline was the lowest. Male and female characters' speech F0 baselines did not differ significantly.

4. Discussion

In our research we tried to determine whether listeners distinguish between the different speaking styles of a male voice artist's reading of an audio novel. Perception tests ascertained that listeners did not struggle to identify the narrator's speech or the male and female characters' direct speech.

Stemming from this knowledge, we studied which among eGeMAPS parameters significantly differentiated these three speaking styles from one another. We recognized 38 parameters, of which most (18) were related to voice quality and timbre. There were 11 parameters related

to voice loudness, eight related to pitch, and one related to tempo (see Table 2). Utilizing these eGeMAPS parameters, it is possible to generate rules that allow labelling the best samples in the corpus for each speaking style in a consistent manner. This would reduce the size of training corpora needed for vivid synthesized speech.

Because interpreting (finding of a perceptual equivalent) of all eGeMAPS parameters is neither meaningful nor possible, we have emphasized some parameters which similar studies have focused on. One such parameter is pitch. Earlier studies have shown that reading out with an imitation of a feminine voice might be done by raising the pitch, and an imitation of a masculine voice might be done by lowering the pitch (Cartei et al. 2012). Yet a study by Stolarski (2017a) showed that although readings of female characters were accompanied by a rise in pitch, the rise in pitch was not statistically significant for most readers. Neither did Stolarski note a drop in pitch for readers portraying male characters' speech. Our study showed that pitch was a significant differentiator between speaking styles, as narrator speech had the lowest and female character speech the highest pitch (see Figure 1). Based on the results, we can also say that direct speech was read out at a higher pitch than the narrator's speech. The same can be ascertained based on the F0 baseline, which was lower for narrator speech in comparison to direct speech (see Figure 7). The pitch of different text structures was also analysed by Mihkla et al. (2018). Their study showed no significant difference in the F0 baseline between direct speech and text adjacent to it. Yet a study by Mihkla et al. (2017) on the agreeableness of the prosodic markers of the text structure as read out by a synthesized voice revealed that listeners preferred direct speech that was read at a higher pitch.

Our results on pitch variability corroborated the results of Stolarski (2017a): male and female characters' speech did not present a significant difference in pitch variability. In addition, our study showed that pitch varied more in narrator speech than in either male or female characters' speech (see Figure 2). This result differs from a comparison of direct and indirect speech done by Yao and Scheepers (2015), where the pitch of direct speech was more variable.

Our results on speech intensity (loudness) differed from those of Stolarski (2017b). That study did not reveal a difference in the intensity of readings of male and female characters' speech, but our study showed that male characters' speech was read out at a significantly louder

volume than female characters' speech. Compared to the narrator's speech, our study found direct speech to be louder (see Figure 3). This gives us a sign that the Estonian style of reading audiobooks differs from American English readings in loudness, where direct speech was performed at a lower intensity, that is, less loudly (cf. Stolarski 2017b). In variability of loudness, our results fell in line with Stolarski's (2017b): male and female characters' speech did not differ in loudness variability. Comparing the narrator's speech and direct speech, our results showed that loudness varied more in direct speech (see Figure 4).

Studies by Yao and Scheepers (2011, 2015) showed that speech tempo varied more for direct speech than for indirect speech. Our study showed that male and female characters' speech tempo varied less than the narrator's speech tempo (see Figure 6), while speech tempo did not play a role in differentiating speaking styles (see Figure 5).

Although our results show some overlap and some discrepancies with earlier studies, we have to refrain from wider generalizations, as the goals and material analysed are not exactly comparable, and the cultural context is different.

Summarizing the results of this study on pitch, loudness, and tempo, we can compare male and female characters' direct speech as follows:

- Female characters' speech was higher and quieter. The speech tempo varied less.
- Male characters' speech was lower and louder. The speech tempo varied more.

A comparison between the narrator's speech and direct speech showed that:

- The narrator's speech was lower and quieter, and pitch and speech tempo varied more.
- Direct speech was higher and louder, while loudness varied more.

If the vividness of speech is characterized by the variability of pitch, loudness, and tempo, then the results of our study showed that our voice artist read the narrator's speech in a more vivid manner than direct speech. Although direct speech was louder and higher, it was more static in variability of pitch and tempo. Based on the acoustic results garnered from the material we used, we cannot say whether this was the specific reading style of one voice artist or characteristic of a general Estonian reading style. Perception test results, where listeners

easily identified the narrator's speech and direct speech (see Table 2), support the assumption that these may be general techniques used by Estonian readers in performing a narrator's speech and direct speech. In future studies we hope to clarify this further.

5. Conclusion

The study was driven by a need to improve the expressivity of audiobooks read by synthesized voices and to ease the differentiation of speaking styles and characters. For this purpose, an analysis was carried out on what speaking styles a voice artist uses to differentiate between a narrator's speech and male and female characters' direct speech and to see whether the difference is recognizable to listeners. The perception tests showed that listeners were able to identify the three speaking styles. Acoustically the styles were differentiated by 38 eGeMAPS parameters, which allow for choosing more consistent material for inclusion in the corpus for training different styles.

Acknowledgements

This study was supported by the Estonian Ministry of Education and Research (IUT35-1), by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies), and by the National Programme for Estonian Language Technology 2018–2027.

Addresses:
Hille Pajupuu
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
E-mail: hille.pajupuu@eki.ee

Rene Altrov
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
E-mail: rene.altrov@eki.ee

Jaan Pajupuu
Industry62
Toompuiestee 35
10133 Tallinn, Estonia
E-mail: jaan.pajupuu@industry62.com

References

Alain, Pierre, Nelly Barbot, Jonathan Chevelu, Gwenole Lecorve, Damien Lolive, Claude Simon, and Marie Tahon (2017) "The IRISA text-to-speech system for the Blizzard Challenge 2017". *Proceedings of the Blizzard Challenge 2017 Workshop*, Stockholm, Sweden. Available online at <http://www.festvox.org/blizzard/bc2017/IRISA_Blizzard2017.pdf>. Accessed on 12.08.2019.

Cartei, Valentina, Heidi Wind Cowles, and David Reby (2012) "Spontaneous voice gender imitation abilities in adult speakers". *PLoS ONE* 7, 2, e31353.

Chalamandaris, Aimilios, Pirros Tsiakoulis, Sotiris Karabetsos, and Spyros Raptis (2014) "Using audio books for training a text-to-speech system." *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014,* 3076–3080.

Charfuelan, Marcela and Ingmar Steiner (2013) "Expressive speech synthesis in MARY TTS using audiobook data and emotionML", *Proceedings of Interspeech 2013*, 1564–1568.

Chistikov Pavel, Dmitriy Zakharov, and Andrey Talanov (2014) "Improving speech synthesis quality for voices created from an audiobook database". In Andrey Ronzhin, Rodmonga Potapova, and Vlado Delic, eds. *Speech and computer. Proceedings of the 16th International Conference, SPECOM 2014 (Lecture Notes in Computer Science* 8773), 276–283. Cham: Springer.

Elson, David K. and Kathleen R. McKeown (2010) "Automatic attribution of quoted speech in literary narrative". *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10),* 1013–1019.

Eyben, Florian, Sabine Buchholz, Norbert Braunschweiler, Javier Latorre, Vincent Wan, Mark J. F. Gales, and Kate Knill (2012) "Unsupervised clustering of emotion and voice styles for expressive TTS". *Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4009–4012.

Eyben, Florian, Felix Weninger, Florian Groß, and Björn W. Schuller (2013) "Recent developments in openSMILE, the Munich open-source multimedia feature extractor". *Proceedings of the 21st ACM International Conference on Multimedia*, 835–838.

Eyben, Florian, Klaus Scherer, Bjorn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong (2016) "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". *IEEE Transactions on Affective Computing* 7, 2, 190–202.

He, Hua, Denilson Barbosa, and Grzegorz Kondrak (2013) "Identification of speakers in novels". *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1312–1320.

Iosif, Elias and Taniya Mishra (2014) "From speaker identification to affective analysis: a multi-step system for analyzing children's stories". *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL),* 40–49.

Lindh, Jonas and Anders Eriksson (2007) "Robustness of long time measures of fundamental frequency". *Proceedings of Interspeech 2007*, 2025–2028.

Mihkla, Meelis, Indrek Hein, Indrek Kiissel, Artur Räpp, Risto Sirts, and Tanel Valdna (2013) "Subtiitrite helindamine - kas, kuidas, kellele ja milleks?" *Keel ja Kirjandus* 11, 819–828.

Mihkla, Meelis, Indrek Hein, Indrek Kiissel, Artur Räpp, Risto Sirts, and Tanel Valdna (2014) "A system of spoken subtitles for Estonian television". In Andrius Utka, Gintarė Grigonytė, Jurgita Kapočiūtė-Dzikienė, Jurgita Vaičenonienė, eds. *Human language technologies – the Baltic perspective. Frontiers in artificial intelligence and applications 268*, 19–26. IOS Press Ebooks. Available online at <http://ebooks.iospress.nl/volumearticle/37998>. Accessed on 10.08.2019.

Mihkla, Meelis, Indrek Hein, Andrus Hiiepuu, Indrek Kiissel, Raivo Ruusalepp, and Urmas Sinisalu (2017) "Raamat sünnib kuulata". *Keel ja Kirjandus* 2, 114–129.

Mihkla, Meelis, Indrek Hein, and Indrek Kiissel (2018) "Self-reading texts and books". In Kadri Muischnek and Kaili Müürisep, eds. *Human language technologies – the Baltic perspective. Frontiers in artificial intelligence and applications* 307, 79–87. Amsterdam: IOS Press.

Montaño, Raúl, Francesc Alías, and Josep Ferrer (2013) "Prosodic analysis of storytelling discourse modes and narrative situations oriented to text-to-speech synthesis", *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW8)*, 171–176.

Montaño, Raúl and Francesc Alías (2016) "The role of prosody and voice quality in indirect storytelling speech: annotation methodology and expressive categories". *Speech Communication* 85, 8–18.

Montaño, Raúl and Francesc Alías (2017) "The role of prosody and voice quality in indirect storytelling speech: a cross-narrator perspective in four European languages". *Speech Communication* 88, 1–16.

Piits, Liisi (2016, November 9) *Eesti Keele Instituudi kõnesünteesikorpus.* Center of Estonian Language Resources. Available online at https://doi.org/10.15155/3-00-0000-0000-0000-05bdal. Accessed on 19.07.2019.

R Core Team (2017) "R: a language and environment for statistical computing". Available online at <https://www.R-project.org>. Accessed on 04.12.2017.

Sini, Aghilas, Damien Lolive, Gaëlle Vidal, Marie Tahon, and Élisabeth Delais-Roussarie (2018) "SynPaFlex-Corpus: an expressive French audiobooks corpus dedicated to expressive speech synthesis". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* 4289–4296.

Stolarski, Lukasz (2017a) "Rendering of gender when reading fiction aloud". *Linguistica Silesiana* 38, 249–283.

Stolarski, Lukasz (2017b) "Intensity of the reader's voice in the reading aloud of fiction: effects of the character's gender". *Studia Anglica Posnaniensia* 52, 3, 285–323.

Yao, Bo and Christoph Scheepers (2011) "Contextual modulation of reading rate for direct versus indirect speech quotations". *Cognition* 121, 3, 447–453.

Yao, Bo and Christoph Scheepers (2015) "Inner voice experiences during processing of direct and indirect speech". In Lyn Frazier and Edward Gibson, eds. *Explicit and implicit prosody in sentence processing: studies in honor of Janet Dean Fodo*, 287–307. (Studies in Theoretical Psycholinguistics*,* 46.) Springer Nature.

Zhang, Jason Y., Alan W. Black, and Richard Sproat (2003) "Identifying speakers in children's stories for speech synthesis". *Proceedings of EuroSpeech,* 2041–2044.

Zhao, Yong, Di Peng, Lijuan Wang, Min Chu, Yining Chen, Peng Yu, and Jun Guo (2006) "Constructing stylistic synthesis databases from audio books". *Proceedings of Interspeech 2006*, 1750–1753.

Kokkuvõte. Hille Pajupuu, Rene Altrov ja Jaan Pajupuu: Teel audioraamatute sünteeskõne elavdamisele. Uurimuse eesmärk oli teada saada, millised olulisemad akustilised parameetrid eristavad audioraamatu lugeja hääles jutustaja kõnet ning mees- ja naistegelaste otsekõnet. Uurimuse tingis vajadus parandada sünteeshäälega loetavate juturaamatute väljendusrikkust ja kõnestiilide eristatavust. Uurimismaterjalina kasutati professionaalse meeshäälega loetud audioromaani „Tõde ja õigus I" põhjal loodud korpust. Et teada saada, kas audioraamatu lugeja hääle põhjal on kuulaja võimeline eristama eri kõnestiile (jutustaja kõnet, mees- ja naistegelaste otsekõnet), koostati 48 lausest koosnev tajutest. Testi tulemused näitasid, et kuulajad tundsid ära kõik kolm kõnestiili. Akustiliseks analüüsiks kasutati kogu korpuse materjali. openSMILE'i tööriistaga ekstraheeriti kõnest iga lause jaoks 88 eGeMAPSis defineeritud parameetrit. Statistiliselt oluliselt eristasid kõnestiile 38 parameetrit, millest 18 oli seotud hääle kvaliteedi ja tämbriga, 11 hääle valjusega, 8 hääle kõrgusega ja 1 tempoga. Kuna tajutest ja akustiliste parameetrite analüüs näitasid, et audioraamatus eristusid nii jutustaja kõne, naistegelaste otsekõne kui ka meestegelaste otsekõne, võib pidada otstarbekaks õpetada juturaamatuid ettelugevaid süntesaatoreid esitama kõiki kolme kõnestiili.

Märksõnad: audioraamatud, kõnestiil, otsekõne, karakteri kõne, GeMAPS, kõneanalüüs, ekspressiivne kõnesüntees