Self-perceived preferences of voice and speaking style characteristics in spoken text

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

[1,2]Swedish Agency for Accessible Media, MTM, Malmö, Sweden
[2]KTH Speech, Music and Hearing, Sweden

christina.tannander@mtm.se edlund@speech.kth.se

Abstract

119 respondents expressed their opinions in a survey on voice and speaking style characteristics in a listening experience context as a step towards a better understanding of which characteristics make for a good voice. We found consensus on some characteristics (e.g. a soft voice is positive, and a forced voice is negative), but also noted that opinion seems conditioned on text type (e.g. dramatic reading is preferred by some fiction listeners but disliked by university textbook listeners).

1 Introduction

We present a survey study of what *voice* and *speaking style characteristics* people associate with the quality of a listening experience, mainly in terms of the ability to maintain focus/concentration. The complete details of the survey is presented in Swedish in MTM (2020). The survey involved no listening but was a straightforward questionnaire.

The study is part of a broader effort to investigate the relation between (1) subjective judgements of spoken text, (2) objective acoustic and prosodic characteristics of spoken text, and (3) efficiency of spoken text in, for example, study situations. A long-term goal is to provide more objective and efficient tools to select voice talents for spoken text production. By *spoken text*, we mean writing that is read aloud, for example audiobooks and talking books (for a discussion of its relation to other representations of language, see Tånnander & Edlund, 2019). Spoken text is one of the adapted formats (alongside e.g. Braille) used to increase accessibility to writing.

The mission of the governmental authority *Swedish Agency for Accessible Media* (MTM) includes providing large quantities of spoken text for people with low vision or reading difficulties. The agency produces fiction (mostly narrated by humans), university text books (>50% with synthetic speech) and over 100 newspapers (all synthetic speech; Tånnander, 2018) .

Hollien et al. (1991) explored preferences on speaker gender, pitch and intensity of younger and older listeners, and found little difference between the groups. 80 voices were judged, and in the top 20 they found female and male voices, low, medium and high f0, and low and mid intensity. They propose three explanations for the lack of results: 1. voice preferences are established in the early childhood and don't change; (2) preferences are based on stereotypes; (3) listeners attended to other vocal characteristics than the ones examined (e.g. intonation patterns or general voice quality).

Goy et al., (2016) investigated pleasantness, naturalness, clarity, ease of understanding, loudness, and the talker's suitability to be an audiobook reader, also for younger and older listeners. The intra-class correlation (ICC) for *suitable to be an audiobook reader* was 0,86 (younger) and 0,77 (older). Overall, they noted a higher ICC for younger listeners, so older listeners showed more diversity. The correlation between *clarity* and *ease* is strong, and there is a significant correlation between *pleasant* and *suitable to be an audiobook reader*. Younger listeners showed a negative bias against older voices.

Other work focus on more emotional speaker traits, such as kind-cruel, humorous-serious, emotional-non-emotional (e.g. Aronovitch, 1976).

Finally, a number of speech synthesis related studies explore v*oice likeability* (e.g. Burkhardt et al., 2011; Schuller et al., 2012; Weiss & Burkhardt, 2010), in which the general opinions of voices are central, but typically not in the context of prolonged listening to information-rich spoken text.

	Positive		Negative	
	UG	FG	UG	FG
SOFT	20	35	2	0
NEITHER BRIGHT NOR DARK	17	37	0	1
FLOWING	20	22	1	2
MIDDLE AGED	11	26	1	1
FEMALE	12	20	1	0
MALE	13	15	1	0
DARK	10	18	3	4
STRONG	8	5	4	8
YOUNG	7	3	3	7
BREATHY	1	6	3	5
OLD	2	6	4	6
BRIGHT	3	5	9	15
DRONING	2	1	19	38
HOARSE	2	2	27	35
CREAKY	0	0	28	44
SHRILL	0	0	31	58
NASAL	0	0	32	57
FORCED	0	0	31	61

Table 1. Counts of positive and negative marks by the university group (**UG**) and fiction group (**FG**) for each voice characteristic (**VC**). The shading illustrates the distribution of marks within each group, with positive mark counts in green and negative mark counts in red. The feature list is sorted on the difference between the sum of positive and negative marks for each feature.

	Positive		Negative	
	UG	FG	UG	FG
NORMAL SPEECH RATE	31	59	3	1
CLEAR	33	51	0	0
TRUSTWORTHY	25	38	0	1
INTERESTING	23	36	0	2
COMMITTED	19	37	0	1
EXPRESSIVE	14	35	4	1
NICE	10	20	0	0
KIND	7	20	0	0
OBJECTIVE	16	14	1	3
INTELLIGENT	9	8	0	0
NEUTRAL	10	10	1	8
LIVELY	8	8	7	6
SERIOUS	2	4	1	5
DRAMATIZED	4	11	19	8
FAST SPEECH RATE	10	6	10	22
SLOW SPEECH RATE	8	1	11	27
JOKEFUL	1	3	19	16
SEXY	0	2	26	27
ANGRY	0	1	32	36
AUDIBLE BREATHING	0	1	25	47
STRESSED	1	1	34	48
CARELESS	0	0	32	48
MONOTONE	0	0	28	58
MOUTH SOUNDS	0	0	32	59

Table 2. Counts of positive and negative marks by the university group (**UG**) and fiction group (**FG**) for each speaking style characteristic (**SSC**). For table design description, see table 1.

2 Method

2.1 Survey software

An online questionnaire, *SurveyMonkey*[1], which we know works well with screen readers, was used to find out how voice and speaking style characteristics affect a listener's ability to listen to spoken text for extended periods of time.

2.2 Respondents

The survey was open to anyone, and the link to the service was spread via MTM's web pages, social media and through user organizations related to MTM. The questionnaire contained nine demographic questions (e.g. age and gender). An additional question about the respondent's experience of listening to spoken text was used to split the respondents in two groups: the University Group (**UG**, with experience from using spoken text in studies at university level) and the Fiction Group (**FG**, without such experience).

2.3 Selection of characteristics

18 voice characteristics (**VC**, mainly attributed to the speaker's voice) and 21 speaking style characteristics (**SSC**, mainly attributed to the manner of speaking) were selected as typical by consensus of three experienced voice and speech researchers (see table 1 and 2 respectively). Furthermore, it

[1] https://www.surveymonkey.com/

		% < 10 min	% 11-30 min	% 31-60 min	% A few hours	% > 4 hours
Listen	**UG TTS**	35	20	28	13	5
	UG HUM	12	5	29	44	10
	FG TTS	38	30	13	15	5
	FG HUM	3	10	8	47	32
Judge	**UG**	84	12	2	0	2
	FG	67	21	11	0	2
Get used	**UG TTS**	38	18	18	12	15
	UG HUM	56	15	20	5	5
	FG	61	16	10	10	4

Table 3. How long respondents can **listen** to different voices, and time needed to **judge** and **get used** to a narration. Some time intervals are merged. TTS = synthetic voice, HUM = human voice.

was important to use a terminology describing voice and speaking style characteristics that also laymen are comfortable with. As we can see in table 2, some characteristics are similar to each other (e.g. nice/kind and objective/neutral).

2.4 Tasks

The respondents were presented to the 18 **VC** and asked to mark as many of them as they liked as either a positive or a negative characteristic, according to their own subjective opinion. The two groups of respondents were given different contexts to bear in mind: **UG** was told to think about listening to university textbooks and **FG** to fiction books. The procedure was repeated for the 21 **SSC**.

Both groups were also asked to estimate several time periods: how long they can listen to a book read by a human or a synthetic voice without losing concentration, how long it takes to decide whether a voice works well or not (this question was not split into human or synthetic voice), and how long time it takes to get used to a specific voice and speaking style. All of these were accompanied by a list of time interval choices (some of which have been merged to create the intervals in table 3). They were also asked open questions about characteristics that affects their ability to maintain or lose concentration, and finally, they were given the opportunity to comment in an open text field.

UG respondents were also presented with a list of speech synthesis specific features (see table 4) and asked "When you listen to university textbooks read by a synthetic voice, what is important to you? You can choose several alternatives."

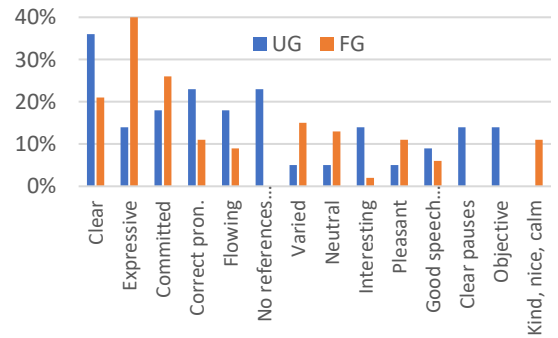

Figure 1. Positive effect on concentration.

3 Results

119 people completed the questionnaire (45 **UG** and 74 **FG**): 67% were women. 47% reported a need for adapted text, for example due to low vision (19%) or reading difficulties (17%).

Table 1 shows how the respondents' marks were distributed for **VC**. On average, **UG** marked 3,2 **VC** as positive and 4,6 as negative. The corresponding numbers for **FG** were 2,9 and 4,7. Table 2 shows the distribution of the respondents' marks for **SSC**. On average, **UG** marked 2,9 **SSC** as positive and 6,4 as negative, and the corresponding **FG** numbers were 3,0 and 5,8.

The responses to the open question about what characteristics affect concentration has been categorized manually, and only the categories that contain more than two respondents from one of the groups are reported. Characteristics that the respondents felt help concentration are in Figure 1, and those that made concentration harder in Figure 2. Table 3 shows the temporal estimates provided

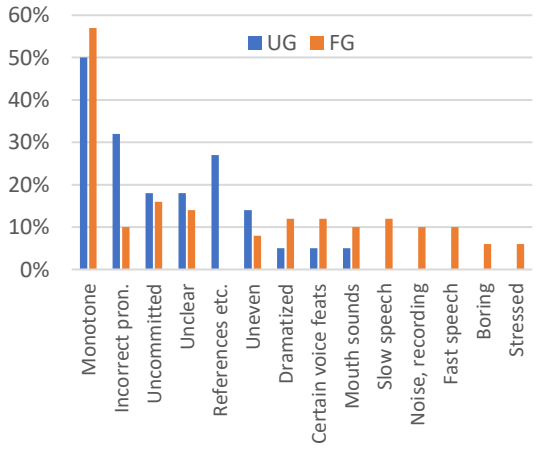

Figure 2. Negative effect on concentration.

	n	%
Clear and easy to understand	37	82
Capable of reading English words in the text	35	78
Possible to skip references and/or footnotes	32	71
All words are pronounced correctly	28	62
Works well listening to at a fast speech rate	21	47
Sounds like a human	21	47
Clear pauses between sentences	20	44
Works good listening to at a slow speech rate	13	29
Clear pauses between phrases	12	27
Other	6	13
Don't know	2	4

Table 4. Important features of a synthetic voice reading university textbooks.

by the respondents in percent. Finally, the responses about speech synthesis features that are important for university textbooks are presented in Table 4.

4 Interpretation

Partly due to sorting of the characteristics, there are few noteworthy differences between the groups at the extremes of the lists in Tables 1 and 2. Numerous characteristics are widely regarded as either positive or negative by both groups. The marks are overwhelmingly positive for the top 7 **VC** and the top 9 **SSC**, and the opposite holds for the bottom 6 **VC** and bottom 8 **SSC**. 5 **VC** (**STRONG**, **YOUNG**, **BREATHY**, **OLD**, **BRIGHT**) and 4 **SSC** (**LIVELY**, **SERIOUS**, **DRAMATIZED**, **FAST SPEECH RATE**) are contentious. We also find clear differences between groups. For example, **UG** shows a stronger preference towards **FAST** and **SLOW** speech and **FG** towards **KIND** speech. Not many respondents prefer **DRAMATIZED** narrations, but the dispreference is high in **UG.** These differences align with the different goals of someone reading fiction or university textbooks. Speaker gender does not matter; some respondents marked female and/or male voices positive but almost no one negative.

Turning to endurance, the most commonly reported maximum time to listen to either human or synthetic voices for **UG** was 31-60 minutes. There is a clear difference in the distribution outside that mode: for human voices, 49% reported lengths of more than one hour, while the corresponding number for a synthetic voice was 16%. In **FG**, 64% reported that they could listen to fiction books narrated by a human voice for more than one hour.

For habituation, the time it takes to get used to a narration, most respondents reported short times. The most common answer in **UG** was up to 10 minutes (56% for human voices and 38% for synthetic). For **FG**, 61% said <10 minutes.

Next, we see considerable consensus between the groups regarding which features help and disturb concentration (figure 1 and 2). A predictable difference is that while both groups rate both clarity and expressiveness highly, the former is of top importance to **UG** and the latter to **FG**.

Figure 3 shows a similar pattern of consensus. Noteworthy exceptions are incorrect pronunciation, which is more disturbing to **UG** than to **FG**, and the reading aloud of references and footnotes, which is only a problem to **UG** (easily explained since the phenomena rarely occur in fiction).

5 Conclusion

There is a large degree of consensus on characteristics perceived as good and bad. However, the fact that people agree in their *self-reported* judgements does not automatically mean that these judgements are good for *objective* measures. Nevertheless, the list of positive and negative characteristics obtained is a fair starting point for further study.

A difference is seen in self-perceived habituation time for synthetic voices and human voices, with the former being longer. This may be alleviated by the fact that there are much fewer synthetic voices to get used to than human voices.

Another, more important conclusion concerns the type of material being read. Our results suggest that different characteristics are important to consumers of spoken text: university textbook readers care about reading speed, the omission of footnotes and references, and the correct pronunciation of foreign and unusual words, while fiction readers care about expressiveness.

Acknowledgements

The surveys were 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

Charles D. Aronovitch. 1976. The Voice of Personality: Stereotyped Judgments and their Relation to Voice Quality and Sex of Speaker. The Journal of Social Psychology, 99(2), 207–220.

Felix Burkhardt, Björn Schuller, Benjamin Weiss and Felix Weninger. 2011. "Would You Buy A Car From Me?" - On the Likability of Telephone Voices. In *Interspeech*. Florence, Italy.

Huiwen Goy, M. Kathleen Pichora-Fuller & Pascal van Lieshout. 2016. Effects of age on speech and voice quality ratings. The Journal of the Acoustical Society of America, 139(1648).

Harry Hollien, Marylou Pausewang Gelfer & Terry Carlson. 1991. Listening Preferences for Voice Types as a Function of Age. Journal of Communication Disorders, 24(2), 157–171.

MTM. 2020. *Webbenkät om att lyssna på uppläst text.* Stockholm, Sweden.

Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi and Benjamin Weiss. 2012. The INTERSPEECH 2012 Speaker Trait Challenge. In *Interspeech*. Portland, OR, USA.

Christina Tånnander. 2018. Speech Synthesis and evaluation at MTM. In *Proceedings of Fonetik*. Gothenburg: Gothenburg University.

Christina Tånnander and Jens Edlund. 2019. First steps towards text profiling for speech synthesis. In *Proc. Digital Humanities in the Nordic Countries 2019 (DHN2019)*. Copenhagen, Denmark.

Benjamin Weiss and Felix Burkhardt. 2010. Voice Attributes Affecting Likability Perception. In *Interspeech*. Makuhari, Japan.