INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Pitch distributions in a very large corpus of spontaneous Finnish speech

Mietta Lennes[1], *Minnaleena Toivola*[1]

[1]University of Helsinki, Finland

mietta.lennes@helsinki.fi, minnaleena.toivola@helsinki.fi

Abstract

Speakers differ in the pitch range they use in their speech. In order to analyze the functional aspects of pitch, the typical pitch range of each individual is needed as reference. However, systematically collected pitch data from a sufficiently large corpus have not been previously available. We analyze the pitch distributions of individual speakers in a subset of the Donate Speech Corpus, collected from speakers of Finnish in 2020–2021. We report pitch analysis results based on samples from 8197 speakers and 1475 hours of speech. We compare the results obtained from male and female speakers in different age groups.

Index Terms: pitch range, voice range, prosody, age-related pitch variation, Finnish

1. Introduction

The pitch range that an individual is able to comfortably produce is largely restricted by the physiological properties of the person in question. People can only control the length and modes of vibration of their vocal folds during phonation up to a certain degree, although the vocal range may be slightly extended by practicing.

In speech, pitch is an important psychoacoustic cue used by the speaker-listeners for chunking and organizing the acoustic speech signal and for monitoring interaction during conversation. The changes in pitch contribute to the perception of prosodic properties such as intonation, stress and accent, but they also serve paralinguistic functions [1]. In tone languages, specific pitch patterns can be used for distinguishing lexical meanings. However, pitch perception is relative: what counts as high or low pitch varies by speaker [2, 3, 4].

In cross-linguistic comparisons, it is common belief that there are differences in pitch variation between languages [5]. However, studies have shown large differences in the methods used to calculate pitch variation as well as in the numbers of speakers. It remains unclear whether the pitch range used in speech would tend to change for a given individual when speaking different languages. It is more likely that the differences between languages tend to occur in the actual pitch patterns used by the speakers.

1.1. The individual pitch range as reference

Previous studies of both speech and singing suggest that individual speakers and singers tend to prefer a certain vocal range. In music, the most comfortable vocal range in singing is referred to as the singer's *tessitura.*

Pitch can be estimated from the speech signal by applying algorithms that aim to detect the fundamental frequency (*f0*). During voiced sounds, the *f0* tends to reflect the glottal frequency in phonation. A general difficulty in pitch detection is that the researcher needs to provide the algorithm with reasonable minimum and maximum pitch values. The accuracy of the detected values tends to suffer in case the boundaries are not optimal for the speaker in question. However, it is difficult to guess the optimal values without manually inspecting the data. This has slowed down the analysis process of many researchers even though a fully accurate pitch analysis is not always required.

In our previous work [6], we showed that the pitch distributions of individual speakers are relatively similar in shape if the measured pitch data are transformed into semitone scale and the speaker-specific statistical pitch mode is used as the reference point. We also presented a method for estimating the typical pitch range by locating the pitch mode in the density function of the pitch data. In an earlier study of Finnish, Russian and Dutch spontaneous speech [7], following a similar approach, no evident differences were observed in the shapes of the mode-referred pitch distributions between speakers of different languages.

According to a study by Moore [8], children and adults tend to favour the lower half of the vocal range that they are able to use in singing. When people are allowed to start singing alone and to choose the key of the song, they tend to select a range where the lowest note of the song is about five semitones above the lower boundary of their vocal range. The highest quarter of the vocal range is rarely used by singers.

1.2. Pitch across the life span

Language-specific studies of the differences in pitch between women and men by age have found evidence that pitch decreases after adolescence for physiological reasons. In addition, the pitch of women's voices starts decreasing by the time of their menopause, whereas men's pitch tends to rise at older age. [9]

In a literature review, Saggio and Constantini [10] reported the *f0* means and standard deviations for homogeneous groups of women and men, aged 20–40 years, from twenty countries. Baseline means showed an *f0* ranging from 202–267 Hz for females and 113–154 Hz for males. The effect of age on *f0* has been assessed in numerous studies. The *f0* in women tends to decrease with age [11, 9, 12], although an increase in *f0* at age 80 has also been observed [13]. There is also uncertainty about age-related changes in *f0* in men. A slight increase in the *f0* of older men has been reported [11, 13, 9], whereas another study reported a decrease in *f0* for men over 60 [14].

In previous studies of the normal voice pitch of both male and female speakers, the results have been partly contradictory,

10.21437/Interspeech.2023-1822

especially regarding age-related changes. The number of participants per age group has often been limited. In some studies, the pitch analysis has been performed on vowel segments in continuous speech, whereas others have measured pitch in sustained vowel productions. There is also variation in the specific methods of selecting and collecting the pitch data points and in the choice of statistical metrics used for reporting the results.

2. Material and Methods

For the present study, we used a subset of the Donate Speech Corpus [15], a very large speech corpus collected during the Donate Speech campaign in Finland starting from June 2020. The corpus contains Finnish speech samples from thousands of volunteers who recorded their speech via a dedicated mobile or desktop app. The speech was elicited under different themes and individual tasks where the speakers were encouraged to talk about a video or image, or specific areas in their daily lives (pets, favourite piece of clothing, feelings under the covid-19 pandemic, etc.). Since the speech recordings mainly consist of monologues, where the speaker is talking to the app and usually alone, the corpus is well suited for studying the speakers' individual preferences. For instance, potential pitch synchronization effects that may occur between speakers [16] can be excluded.

In the Donate Speech app, the speakers were also requested to provide some background information about themselves, including gender, age group, native language, area of residence, dialectal background, education level, profession, etc. When designing the background questions, any potentially identifying details were avoided. The speakers were allowed to skip these questions, and they may also have provided false details. Therefore, the speaker metadata are only partial and not fully reliable. For a more detailed description of the Donate Speech campaign and the design principles of the corpus, see Lindén et al. [17].

The first version of the complete Donate Speech Corpus [15] contains about 3200 hours of spontaneous speech. For further inspection, we selected the recordings in the data packages 1–12 from the year 2020. These packages contained a total of about 1475 hours of speech from 8390 different client ID's, which we consider as roughly corresponding to the number of individual speakers. However, several speakers may have donated their speech by using the same mobile device or computer, and thus the number should not be considered as an exact one.

In all the audio files and within entire recordings, pitch was analyzed at 20 ms intervals by using a script written for Praat [18], applying the default autocorrelation method for pitch detection [19] in the Praat program. During the first analysis pass, the minimum pitch parameter was set at 50 Hz and the maximum pitch at 600 Hz, the rest of the analysis parameters remaining at their default values. These settings would hardly be ideal for any individual speaker, since the algorithm would probably tend to smooth out fast pitch changes and it might also pick pitch candidates that are an octave too high. However, this method was used in order to try and ”bootstrap” the speaker-specific modal pitch range, i.e., the span within which a majority of pitch values tend to fall for each speaker [6].

All the defined pitch values were gathered into data files along with the name of the original audio file. The full set of pitch data was then analysed in RStudio [20] and combined with the background details that had been provided by the speakers at the time of their speech donations.

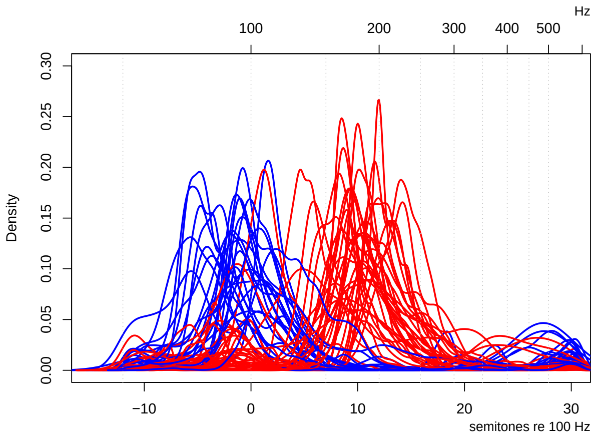

Figure 1: *The pitch distributions of female (red) and male (blue) speakers (N=60) in spoken Finnish. Pitch was detected by applying the pitch floor frequency of 50 Hz and the ceiling frequency of 600 Hz for all speakers.*

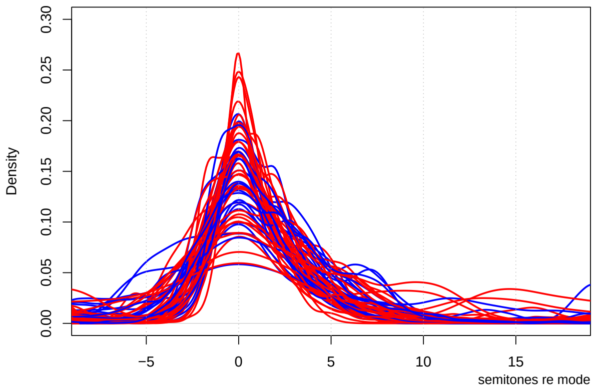

Figure 2: *Distributions of pitch values as referred to the speaker-specific pitch modes of female (red) and male (blue) speakers (N=60). The same pitch floor (50 Hz) and ceiling (600 Hz) frequencies were used in pitch detection for all speakers.*

The density functions of the pitch distributions resulting from the first analysis pass for a sample of 60 individual speakers are plotted in figure 1, using the pitch values recorded in semitone scale with reference to the frequency of 100 Hz. It is seen that the pitch distributions of most, but not all, female speakers tend to be located above those of most male speakers.

In figure 2, the distributions of the same 60 speakers are plotted by using the individual pitch modes as the reference point, located at 0 ST. The individual pitch modes were determined as the pitch at the highest peak of the density function of the pitch values. These examples illustrate the fact that the detected pitch values tend to accumulate within a range of similar size around the pitch mode. Smaller, secondary peaks may occur outside the primary pitch range. Some of these pitch values are likely to be artifacts of the generic analysis parameters that were blindly applied on all types of voices.

In the second analysis pass, the same audio files were analysed again, this time by using the speaker-specific low

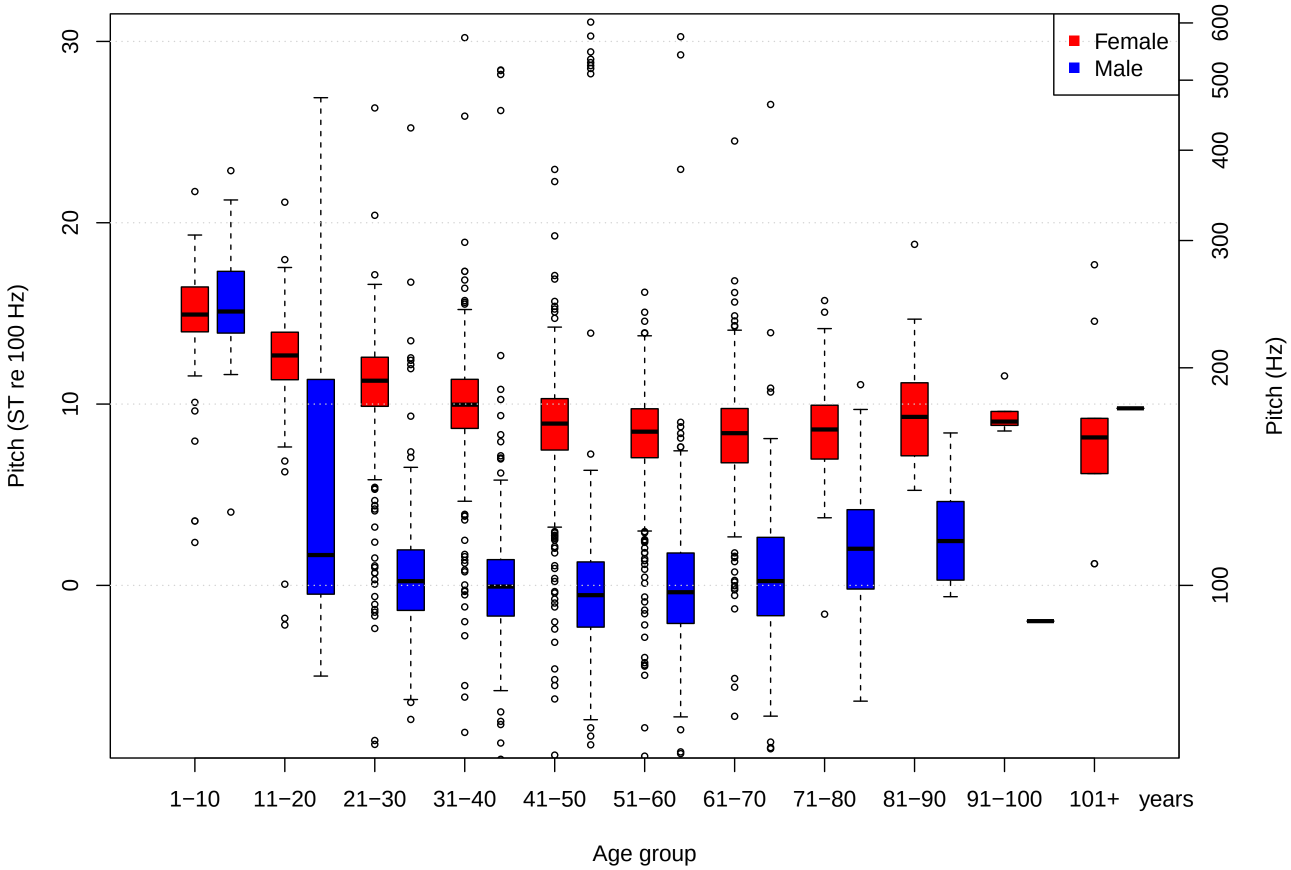

Figure 3: *The distributions of the individually calculated pitch modes of female and male speakers.*

and high boundaries of the modal range as the pitch floor and ceiling parameters (again, keeping the rest of the parameters at their default values). For all speakers, the pitch floor and ceiling values were set at five semitones below each individual's pitch mode and at ten semitones above the mode. The goal of the second analysis pass was to reach a higher resolution of the pitch values analyzed within the speaker's typical speaking voice range. However, for the purposes of the present study, the quality of the results was not assessed in detail, and it is possible that the principles for automatically setting the individual analysis parameters could be improved.

Due to the original data collection method of the Donate Speech Corpus, some client ID's were supplied with some contradictory background details. Since it was impossible to ascertain whether several speakers might have donated under the same ID, we decided to exclude the problematic client ID's from further comparison. We only selected those client ID's that had provided their gender as either ”male” or ”female”, i.e., excluding empty and contradictory answers as well as the options ”other” or ”do not wish to answer”. We also excluded client ID's that were associated with multiple age groups, and client ID's with either none or contradictory information about native language. In order to ensure reliable pitch distributions, we also excluded client ID's with less than 300 pitch values. Finally, we were left with pitch data from 8197 unique client ID's, which we will refer to as ”speakers”.

3. Results

A boxplot of the individually calculated pitch modes of the female and male speakers in each age group is shown in figure 3. For further comparison, table 1 provides the medians of the individual pitch modes and means, expressed in Hertz (Hz) and semitone (ST) scales, and the medians of individual standard deviations of pitch in ST, grouped by self-reported gender and age.

The pitch modes of the male speakers in the age bracket of 11–20 years exhibit a great deal of variation, which probably reflects the fact that speakers undergo their puberty at those ages. The typical pitch of female speakers apparently does not change as radically as the pitch of males after childhood, although the figure suggests a downward trend for women until they reach their fifties or sixties. For the female speakers in this data set, the lowest pitch modes were observed in the age group of 61–70 years.

The changes with age are also broadly in line with previous studies that have found a lowering of voice pitch with age, but our results suggest more specific tendencies for typical pitch changes in certain age groups. For men over 20 years of age, the lowest typical pitches apparently occur in the 41–50 age group, although the differences with the previous age group or with the next one are not large. The pitch modes for male speakers exhibit a gradual increase in the elder age groups.

The differences between the medians of the typical pitches of male and female speakers are particularly interesting.

Table 1: *Medians of the speaker-specific means, standard deviations and statistical modes of voice pitch, grouped by self-reported gender and age. The medians of the means and modes are expressed in both Hertz scale and in semitones with reference to 100 Hz. The standard deviations were calculated from the original pitch values in semitones. The rightmost column shows the difference in semitones between the medians of the pitch modes of the female vs. male groups.*

Age	**N**		**Mean (Hz)**		**Mean (ST)**		**Stdev (ST)**		**Mode (Hz)**		**Mode (ST)**		**Diff. (ST)**
group	Female	Male	F	M	F	M	F	M	F	M	F	M	F-M (mode)
1-10	69	43	264.4	252.7	16.8	16.0	3.1	3.1	237.0	239.3	14.9	15.1	−0.2
11-20	386	137	221.7	119.0	13.8	3.0	2.4	2.7	208.0	110.2	12.7	1.7	11.0
21-30	1348	414	204.6	106.5	12.4	1.1	2.5	2.6	191.9	101.3	11.3	0.2	11.1
31-40	1155	421	189.4	105.4	11.1	0.9	2.7	2.7	177.8	99.6	10.0	−0.1	10.0
41-50	960	357	179.3	103.4	10.1	0.6	2.7	2.7	167.4	96.9	8.9	−0.5	9.5
51-60	1150	472	174.8	103.6	9.7	0.6	2.8	2.7	163.2	97.8	8.5	−0.4	8.9
61-70	613	321	173.7	106.4	9.6	1.1	2.8	2.6	162.4	101.4	8.4	0.2	8.2
71-80	197	124	173.3	116.8	9.5	2.7	2.7	2.5	164.3	112.4	8.6	2.0	6.6
81-90	21	20	174.1	124.3	9.6	3.8	2.8	2.8	171.0	115.2	9.3	2.4	6.8
91-100	5	1	186.9	93.9	10.8	−1.1	2.4	2.1	168.5	89.2	9.0	−2.0	11.0
101+	9	1	162.8	190.7	8.4	11.2	2.8	2.7	160.2	175.8	8.2	9.8	−1.6
All ages	5913	2311	187.6	106.4	10.9	1.1	2.7	2.7	175.2	100.9	9.7	0.2	9.6

As seen in table 1, the pitch modes of the two genders are the most distinct in the age of 21–30 years. For the young adults, the difference between the medians of modes is approximately 11 semitones, i.e., slightly less than an octave. The difference in pitch modes between adult male and female speakers is the least pronounced at ages of 71–80 and 81–90 years. However, conclusions cannot be reliably drawn with respect to the eldest speakers, however, since the numbers of speakers representing these age groups are very low in this data set.

Our data show both similarities and differences with previous research. For both female and male adult speakers, the changes in the individual typical pitch used in speech tend to follow a somewhat similar pattern: lowering from childhood to adulthood, and then perhaps slightly rising after middle age. According to this data set, male voices tend to be at their lowest around 41-50 years, whereas female voices are lowest around the age of 61-70 years. However, it is to be noted that the current data are not longitudinal but cross-sectional, i.e., trends across the life span of individuals cannot be established. However, the data do provide some reference points of the pitch levels that tend to be preferred by speakers of different ages.

The pitch means of male and female speakers in the age group of 20–40 years appear to be slightly lower than those reported in the study by Saggio et al. [10]. It is possible that Finnish speakers tend to have lower-pitched voices than speakers of other languages. However, since the pitch mean can be very sensitive to extreme pitch values and there are differences between the pitch analysis procedures applied in previous work, the result needs to be confirmed.

There are some general limitations to the pitch detection method that may result in individual erroneous values, for instance in the case of creaky voice where regular periodicity might not be evident in the speech signal. Thus, the pitch distributions resulting from the present analysis workflow may not be easy to interpret for all speakers. Moreover, there may be some issues related to the metadata erroneously reported by the speakers, and it is even possible that some of the recordings in the data set do not contain speech at all. Nevertheless, we believe that the majority of the data are representative of the speaker groups under investigation.

4. Conclusions

The Donate Speech Corpus is an exceptionally large speech corpus of spontaneous speech that can provide a cross-section of the phonetic properties of spoken Finnish. In the present study, speaker-specific pitch distributions were analyzed in a 1475-hour subset of the donated speech data covering nearly 8200 individual speakers of Finnish. We applied a two-pass pitch analysis procedure where the minimum and maximum limits of the modal pitch range exhibited by individual speakers were first estimated from the data, and a second pitch detection pass was then performed by using the speaker-specific analysis parameters.

The results suggest that it is possible to automatically estimate the most typical pitch for an individual speaker without knowing in advance whether the speaker is male, female, or a child. Although some speaker-specific variation does occur in the exact shape of the modal pitch distribution, it is suggested that the most comfortable pitch range for each individual tends to be distributed between about five semitones below the pitch mode and about ten semitones above the mode. Tentatively, it is suggested that this is also the most likely range within which the more detailed linguistically and interactionally relevant pitch variation typically takes place.

By using the background information provided by the speech donors in the corpus, it was also shown that adult male and female speakers of different ages tend to differ in the most typical pitch of their speech. In this data set, the typical pitch difference between male and female speakers tends to be at its largest after teenage and for young adults. However, longitudinal studies are needed in order to confirm in more detail when and how pitch changes may take place during the life spans of individual speakers.

5. References

[1] D. R. Ladd, *Intonational phonology.* Cambridge: Cambridge University Press, 1996.

[2] J. Leather, "Speaker normalization in perception of lexical tone," *Journal of Phonetics*, vol. 11, pp. 373–382, 1983.

[3] C. B. Moore and A. Jongman, "Speaker normalization in the perception of mandarin chinese tones," *The Journal of the Acoustical Society of America*, vol. 102, pp. 1864–1877, 1997.

[4] E. Couper-Kuhlen, "The prosody of repetition. on quoting and mimicry," in *Prosody in Conversation*, E. Couper-Kuhlen and M. Selting, Eds. Cambridge: Cambridge University Press, 1996.

[5] I. Mennen, "Second language acquisition of pitch range in german learners of english," *Studies in Second Language Acquisition*, vol. 36, no. 2, pp. 303–329, 2014. [Online]. Available: https://www.jstor.org/stable/26328942

[6] M. Lennes, M. Stevanovic, D. Aalto, and P. Palo, "Comparing pitch distributions using Praat and R," *Phonetician*, no. 111-112, pp. 35–53, 2015.

[7] M. Lennes, D. Aalto, and P. Palo, "Puheen perustaajuusjakaumat: Alustavia tuloksia," in *Fonetiikan päivät 2008. XXV Fonetiikan päivillä Tampereen yliopistossa 11.-12.1.2008 pidetyt esitelmät*, ser. Tampere Studies in Language, Translation and Culture, Series B 3, M. O'Dell and T. Nieminen, Eds. Tampere: Tampere University Press, 2009, pp. 147–155. [Online]. Available: https://urn.fi/urn:isbn:978-951-44-7580-1

[8] R. S. Moore, "Comparison of children's and adults' vocal ranges and preferred tessituras in singing familiar songs," *Bulletin of the Council for Research in Music Education*, vol. 107, pp. 13–22, 1991. [Online]. Available: http://www.jstor.org/stable/40318417

[9] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *Journal of Voice*, vol. 32, no. 5, pp. 644.e1–644.e9, 2018.

[10] G. Saggio and G. Costantini, "Worldwide healthy adult voice baseline parameters: a comprehensive review," *Journal of Voice*, vol. 36, no. 5, pp. 637–649, 2022.

[11] M. Nishio and S. Niimi, "Changes in speaking fundamental frequency characteristics with aging," *Folia Phoniatr Logop*, vol. 60, pp. 120–127, 2008.

[12] L. Albuquerque, C. Oliveira, A. Teixeira, P. Sa-Couto, and D. Figueiredo, "A comprehensive analysis of age and gender effects in european portuguese oral vowels," *Journal of Voice*, vol. 37, no. 1, pp. 143.e13–143.e29, 2023.

[13] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011.

[14] S. Deliyski and D. A. Xue, "Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications," *Educational gerontology*, vol. 27, no. 2, pp. 159–168, 2001.

[15] University of Helsinki, "Donate Speech Corpus, version 1.0," 2022. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2020090321

[16] S. Amiriparian, J. Han, M. Schmitt, A. Baird, A. Mallol-Ragolta, M. Milling, M. Gerczuk, and B. Schuller, "Synchronization in interpersonal speech," *Front Robot AI*, vol. 6, no. 116, 2019.

[17] K. Lindén, T. Jauhiainen, M. Lennes, M. Kurimo, A. Rossi, T. Kurki, and O. Pitkänen, "Donate speech: Collecting and sharing a large-scale speech database for social sciences, humanities and artificial intelligence research and innovation," in *CLARIN : The Infrastructure for Language Resources*, A. W. Darja Fišer, Ed. Berlin: de Gruyter, 2022, pp. 481–510.

[18] P. Boersma and D. Weenink. (2022) Praat: doing phonetics by computer (Version 6.3.02). [Computer program]. Available: https://www.praat.org/. Retrieved on 29.11.2022.

[19] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[20] Posit Software, PBC, "RStudio 2022.12.0 build 353," [Computer program], 2022, available: https://posit.co/downloads/.