Comparing pitch distributions using Praat and R

Lennes, Mietta

2015

Lennes , M , Stevanovic , M , Aalto , D & Palo , P 2015 , ' Comparing pitch distributions using Praat and R ' , Phonetician , no. 111-112 , pp. 35-53 . < http://www.isphs.org/Phonetician/Phonetician_111-112.pdf >

http://hdl.handle.net/10138/351051

COMPARING PITCH DISTRIBUTIONS USING PRAAT AND R

Mietta Lennes[1], Melisa Stevanovic[2], Daniel Aalto[3], and Pertti Palo[4]
[1] Department of Modern Languages, University of Helsinki, Finland
[2] Finnish Centre of Excellence on Intersubjectivity in Interaction, University of Helsinki, Finland
[3] Rehabilitation Medicine, Communication Sciences and Disorders, University of Alberta, and Institute for reconstructive sciences in medicine, Misericordia Community Hospital, Edmonton, Canada
[4] Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, United Kingdom
e-mail: mietta.lennes@helsinki.fi, melisa.stevanovic@helsinki.fi, aalto@ualberta.ca, pertti.palo@gmail.com

Abstract

Pitch analysis tools are used widely in order to measure and to visualize the melodic aspects of speech. The resulting pitch contours can serve various research interests linked with speech prosody, such as intonational phonology, interaction in conversation, emotion analysis, language learning and singing. Due to physiological differences and individual habits, speakers tend to differ in their typical pitch ranges. As a consequence, pitch analysis results are not always easy to interpret and to compare among speakers.

In this study, we use the Praat program (Boersma & Weenink 2015) for analyzing pitch in samples of conversational Finnish speech and we use the R statistical programming environment (R Core Team, 2014) for further analysis and visualization. We first describe the general shapes of the speaker-specific pitch distributions and see whether and how the distributions vary between individuals. A bootstrapping method is applied to discover the minimal amount of speech that is necessary in order to reliably determine the pitch mean, median and mode for an individual speaker. The scripts and code written for the Praat program and for the R statistical programming environment are made available under an open license for experimenting with other speech samples. The datasets produced with the Praat script will also be made available for further studies.

1 Introduction

The analysis of the melodic aspects of speech serves various research interests, such as intonational phonology, speech communication, interactional linguistics, interactional sociology, emotion analysis and language learning. Relative pitch levels and patterns can be connected with many language-specific linguistic functions, such as intonation, stress, (sentence) accent or lexical tones. In conversation, subtle

variations in pitch have been shown to convey, for example, turn-taking or turn-yielding intentions (Duncan, 1972; Ford & Thompson, 1996; Szczepek-Reed, 2004), sequence organization (Kaimaki, 2010; Persson 2013), information status (Breen et al., 2010) and confidence (Scherer et al., 1973).

The pitch range of a speaker depends on physiological (Titze, 1989) and psycho-social (e.g., Cartei et al., 2014; Munson et al., 2015) factors and can serve as an identifying characteristic of the speaker (Kinoshita et al., 2009; Munson, 2007). Due to this variability, theories of intonational phonology usually work with relative pitch levels or excursions within utterances (see, e.g., Ladd, 1996 for a detailed discussion) and not absolute pitch. Moreover, the functional significance of pitch in conversation depends not only on its absolute levels but largely on its relation to the speaker-specific pitch range (e.g., Couper-Kuhlen, 1996). In other words, what counts as high or low varies by speaker (Leather, 1983; Moore & Jongman, 1997). These insights are supported by empirical research showing that listeners are capable of locating the pitch of a given speech sound within the speaker's range without external context or previous exposure to the speaker's voice (Honorof & Whalen, 2005). Thus, in order to analyze the pitch of a given speaker, it is necessary to relate it to his or her typical pitch range.

Since the present study deals with perceptual and relative properties in speech, we prefer to use the term *pitch* instead of the acoustic concept of fundamental frequency (f_0) in this work. The choice of scale plays an important role in analyzing pitch variation. Fundamental frequency f_0, which correlates non-linearly with the perceived pitch in voiced speech, is measured and reported as absolute values in Hertz scale. Traunmüller and Eriksson (1995) provide an overview of previous reports concerning the f_0 ranges of male and female speakers. They point out that when the f_0 range is expressed in the absolute Hertz scale, female speakers appear to exhibit a wider range than men, but the difference more or less disappears when the data are converted into semitones. When expressed in semitone scale, the overall shapes of pitch distributions appear to be similar between speakers (Lennes, 2007) and even between different languages (Lennes et al., 2008). This is not surprising, since humans have similar vocal organs, and the vocal folds can only be stretched within certain limits. Moreover, during modal phonation, it is not possible to instantaneously jump from low pitch to high pitch or vice versa, but the speaker will have to glide through the intermediate pitch levels.

The aim of the present work is to investigate the general distribution of pitch in conversational Finnish speech and to discover the minimum requirements for obtaining reliable statistics of speaker-specific pitch ranges. We will first calculate and describe the pitch distributions of 40 Finnish speakers in everyday conversation, pinpointing some factors that may affect the typical distribution shape in individual cases. Using a bootstrapping method, we will then attempt to determine the minimum amount of samples that is required in order to calculate the mean, median or mode.

We invite other researchers to replicate the results and to extend and improve the method. For these purposes, the code for Praat and R, as well as the pitch data produced for this study, will be shared online under an open license. Our actual workflow is described more explicitly in the documentation of the scripts. Since the

tools may be of interest to readers without a background in phonetics, we will first briefly describe how human speakers may vary in their preferred pitch ranges and how automatic pitch analysis generally works.

2 Background

Speakers tend to differ in the pitch region they usually employ during speech. This variability in preferred pitch is partly due to anatomical and physiological differences. On average, men have longer and thicker vocal folds than women (e.g., Titze, 1989). This is largely why female speakers tend to speak at a higher pitch than male speakers. Similarly, small children tend to use a much higher pitch region than adults.

In addition to the aforementioned physical differences, people also exhibit culture-specific and idiosyncratic ways of using their voice while speaking or singing. Some speakers may be perceived to have "lively" voices, whereas others may sound "monotonous". This may mean that some speakers employ larger pitch ranges, whereas others prefer to keep their pitch close to their personal level of comfort. On the other hand, some speakers creak almost all the time, whereas others use a breathy voice quality or one that may sound like falsetto. In various medical conditions or as a consequence of a surgical treatment affecting the upper airways, the pitch of a person's voice may change significantly. All in all, voice and pitch are an important part of a person's self and identity.

Since people are apparently able to estimate the general height of each others' voices almost instantly, it is likely that this impression is not based on, e.g., the highest and lowest pitches, which would vary from one utterance and situation to the next. Instead, listeners are more likely to "tune in" to the pitch region that the speaker uses most of the time. In music, the typical, most comfortable pitch range of a singer is sometimes referred to as the *tessitura*.

Thus, in order to be able to compare speakers reliably, it is necessary to determine the typical or preferred pitch range of a particular speaker. However, this is not a technically straightforward task. The automatic analysis of pitch or fundamental frequency in speech does not always provide data that can be easily interpreted and compared among speakers. In addition, poor technical quality of the speech material can distort the analysis result. In order to get plausible data, researchers need to be aware of the general properties and inherent limitations of the pitch extraction algorithm that is being applied.

2.1 Automatic pitch detection

In automatic pitch analysis, the voiced portions of speech are expected to represent a single quasi-periodic sound source. This is true in recordings where only one speaker is speaking at a time and the speaker's vocal folds are vibrating normally and rather steadily. Pitch analysis is usually tuned so as to pick up the fundamental frequency f_0, which usually corresponds to detecting the presence and frequency of the slowest, at least nearly periodic component in the complex acoustic signal. At least during modal (regular) phonation, the f_0 thus reflects the frequency of the glottal pulses, i.e., the repetitive opening-closing sequences of the vocal folds.

There are various methods available for automatic pitch extraction and for representing the resulting pitch contours. In this study, we apply the standard autocorrelation method available in the Praat program (the command **To Pitch...**). This method is often used for studying intonation in speech, whereas the cross-correlation algorithm, also available in Praat, is suited for special purposes, such as voice analysis. In practice, both algorithms calculate a sequence of pitch values using short, partly overlapping time windows or frames extracted from the original audio signal. The resulting values can be plotted as a pitch contour as a function of time, or they may be further analyzed.

Since the larynx and the articulatory organs are rarely held completely steady during speech, the frequency structure of the speech signal changes practically all the time. Each analysis window may include speech that is only partly voiced and/or where the f_0 is changing. In order to be able to select the best or most plausible candidate among a number of all possible pitch candidates within each analysis window, the pitch algorithm requires the user to supply the minimum and maximum frequencies prior to the analysis. These parameters can be adjusted according to the expected frequencies for a particular speaker or for specific analysis purposes. The minimum frequency parameter defines the duration of each analysis window. In order to detect a low f_0, where the glottal periods are relatively long, the analysis window needs to be wider than for a high f_0. However, in case the minimum parameter is set too low, the wide analysis frame will conceal fast changes in the f_0. In addition, users can also adjust more advanced parameters that control the tolerance for abrupt pitch changes between consecutive analysis frames. These parameters are used in the pitch algorithm, since human speakers are not able to shift the pitch of their voice up or down at an arbitrary rate. Nevertheless, it is to be noted that even if all the parameters are set in an appropriate way, external noises and overlapping speakers may distort the result.

Non-modal phonation, such as creaky voice, occurs quite frequently in everyday talk (Ogden, 2001; Gobl & Ní Chasaide, 2003; Yuasa, 2010). Irregular periodicity or two simultaneous glottal modes of vibration may occur during creaky or glottalized phonation, and they are difficult to analyze consistently with the standard pitch algorithms. Such events will often result in missing values, potentially erroneous values with halved or doubled frequency (often referred to as "octave jumps"), or other outliers in the pitch curve. In these cases, it is still possible to perform a partly manual analysis in Praat in order to check the result. This can be accomplished for instance by editing a Pitch object. Alternatively, a PointProcess object can first be generated from the Pitch and the corresponding Sound object. Next, the locations of the automatically detected pitch periods can be edited in the PointProcess editor, after which the PointProcess can be converted back to a Pitch object. Manual editing is applied for instance in the ProsodyPro system, which is intended for the analysis of pitch contours on more large-scale material (Xu, 2013). However, manual work is time-consuming, somewhat subjective, and error-prone. On the other hand, it would be efficient to analyze large amounts of data in batch mode, but even if the pitch analysis parameters are individually adjusted for each speaker, it may not be

ultimately possible to avoid the halved or doubled frequency values. It would be useful to be able to automatically discover which regions of the pitch distribution are likely to represent the speaker's modal voice and which parts are potentially less reliable.

3 Material

We built our analysis on two corpora of conversational speech. The FinDialogue corpus, a part of the larger FinINTAS corpus, contains ten conversations (five male-male dyads and five female-female dyads) between young, native Finnish-speaking adults. The participants in each dialogue knew each other well. The dialogues were recorded in an anechoic room using high-quality headset microphones (AKG HSC-200 SR). The two speakers in each dialogue were sitting a few meters apart and facing opposite directions. They were instructed to chat freely for 45-55 minutes either on a few given topics or on whatever they felt like talking about. Each speaker's voice was recorded with a DAT recorder (Tascam DA-P1) on a separate track in a stereo file and downsampled to a rate of 22050 Hz (sample size 16 bit). Thus, it was possible to analyze each speaker's voice in isolation when required. This corpus will be referred to as Corpus A.

The other collection of conversational Finnish speech, which we shall call Corpus B, consists of shorter dialogues with 8 adult female and 8 adult male speakers (3 male-male dyads, 3 female-female dyads, and 2 male-female dyads). The dialogues were recorded in various conditions using one or two microphones. The dialogues included two mundane telephone conversations (2-3 minutes each), two informal planning interactions in a workplace setting (5 minutes and 20 minutes), and three conversations, where the participants were engaged in a joint decision-making task in an experimental setting (2-4 minutes each). The speakers are referred to with a number preceded by the letter F for female and M for male speakers.

4 Analysis

The analysis procedure of this study was implemented as two main scripts: one for collecting the pitch data from the original audio files in Praat, and the other for running various analyses on the pitch data and for plotting the figures using the R statistical programming environment. The two scripts are available and documented on GitHub (Lennes, 2016).

Using a Praat script (see Lennes, 2016 for a detailed description), all the audio files were analyzed with the standard, autocorrelation-based pitch algorithm available in Praat. The distance between consecutive analysis frames was set to 0.02 seconds, resulting in 50 observed pitch values per second in the measured data.

In a first analysis pass, the default minimum frequency parameter was set at 50 Hz and the maximum at 600 Hz. (The default parameters can be changed in the Praat script for other experiments.) These parameters would be too far apart for almost all adult speakers, i.e., the minimum would be clearly below the lowest fundamental frequency that most male speakers would tend to use, and the maximum value would exceed most of the f_0 values of female speakers. The intention was that these settings

would be likely to create anomalies in the initial pitch data. After this first analysis pass, speaker-specific minimum and maximum frequencies were manually determined by inspecting the pitch distributions in R and by locating and generously delineating the pitch cluster with the highest density in each distribution. The speaker-specific parameters were applied in the second analysis pass so as not to include extremely low or high pitch values.

In total, three different datasets were obtained. Dataset 1 was calculated from raw audio using the default minimum and maximum parameters. This type of analysis can, in principle, be done for any audio file without knowing anything of the speaker(s), although the results will not be reliable. Dataset 2 was produced by applying the speaker-specific pitch parameters to analyze the raw audio. This way, it was possible to see how the pitch distribution was affected by whether the minimum and maximum parameters were set individually or not. In order to save some disk space, all undefined pitch values were excluded from these first two datasets. It should be noted that Datasets 1 and 2 are considered as experimental and they will not be useful for audio files that include more than one speaker. Dataset 3 was calculated from the annotated corpora so that only those parts of the audio signals were analyzed where the speaker in question was actually speaking, according to the utterance-level annotations in the TextGrid files. Dataset 3 was used for comparing speaker-specific distributions.

The frequency values from all the individual analysis frames obtained for all three datasets were automatically written to data tables (tabulated text files) in both Hertz and semitones with respect to the frequency of 100 Hz. For Dataset 3, a total of 489,485 pitch analysis frames, including 277,384 voiced ones, were recorded. A pitch difference expressed in semitones corresponds to the respective musical interval, which makes the data easier to read and interpret. For instance, a difference of 12 semitones (ST) corresponds to an octave, an interval of 7 ST corresponds to a perfect fifth and 5 ST to a perfect fourth. In this article, all pitch values expressed in semitones are provided relative to 100 Hz, unless another reference level or comparison is mentioned.

5 Results

The analysis continued by visualizing the general properties of the pitch distributions for each individual speaker. Since our aim was to estimate the shape of the overall pitch distributions of individual speakers and since pitch and frequency are continuous variables, we first plotted the probability density curves for all speakers and for all three datasets for inspection. A density plot is a continuous version of the more familiar histogram.

5.1 Probability density

The pitch distributions for one female speaker (F3 in Corpus A) are plotted in Figure 1. The analysis calculated from the unannotated audio (Dataset 1) is indicated with a dotted line, Dataset 2 with a dashed line, and Dataset 3 with a solid line. It is observed that the main distribution is skewed to the right. The speaker generally stays around her typical pitch level (mode = 9.8 ST, 176 Hz), but she sometimes goes approximately 6 semitones below or 12 semitones above her mode. Since the audio

signal was of high technical quality, this is probably why there is very little difference between the Dataset 1 distribution, calculated from raw audio with the default parameters, and the result of the more speaker-specific analysis in Dataset 3.

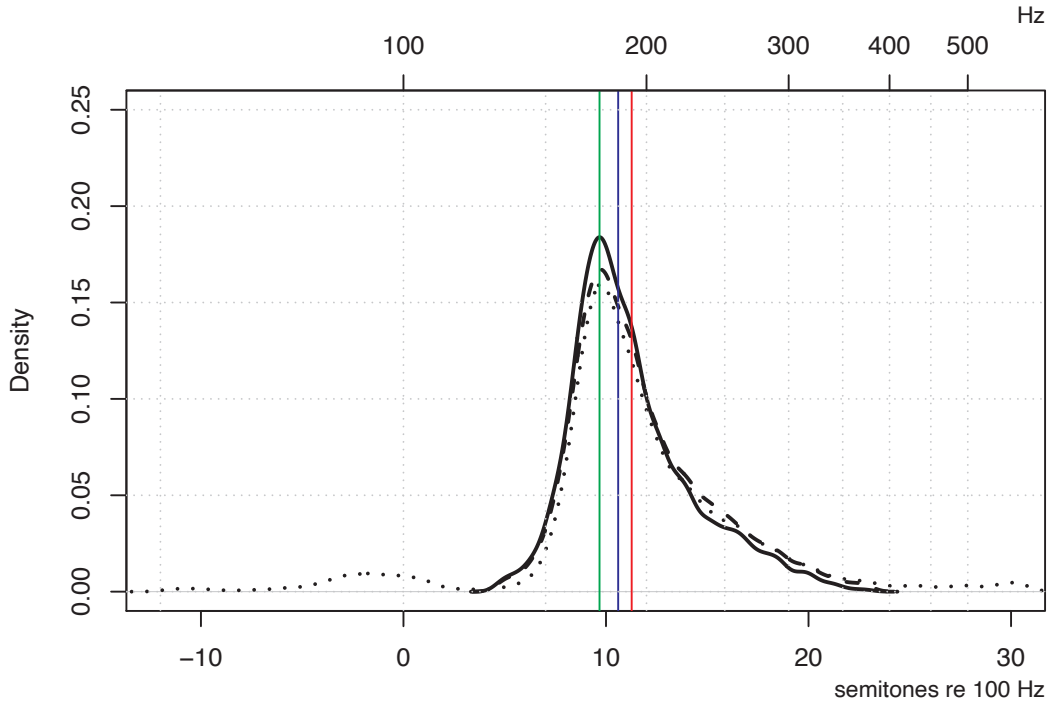

Figure 1: The probability density function of the pitch values obtained from conversational speech recorded from one female speaker (F3 in Corpus A). The mean pitch (11.27 semitones above 100 Hz) is indicated with a red vertical line, median (10.6 ST) with blue and the pitch mode (9.68 ST) with a green line. The corresponding values in the absolute Hertz scale are 195 Hz, 184 Hz and 174 Hz. The dotted line represents the pitch distribution obtained from raw audio (Dataset 1), the dashed line is the distribution calculated from raw audio with manually defined speaker-specific parameters (Dataset 2), and the solid line represents the data calculated within annotated utterances only (Dataset 3).

Another example of the pitch distributions is shown in Figure 2 for the female speaker F23 in Corpus B. In this case, Dataset 1 includes an external low-frequency noise. The total amount of data for this speaker was small (2282 samples in Dataset 3), which is probably the reason why the distribution looks more irregular than that of speaker F3 (12640 samples).

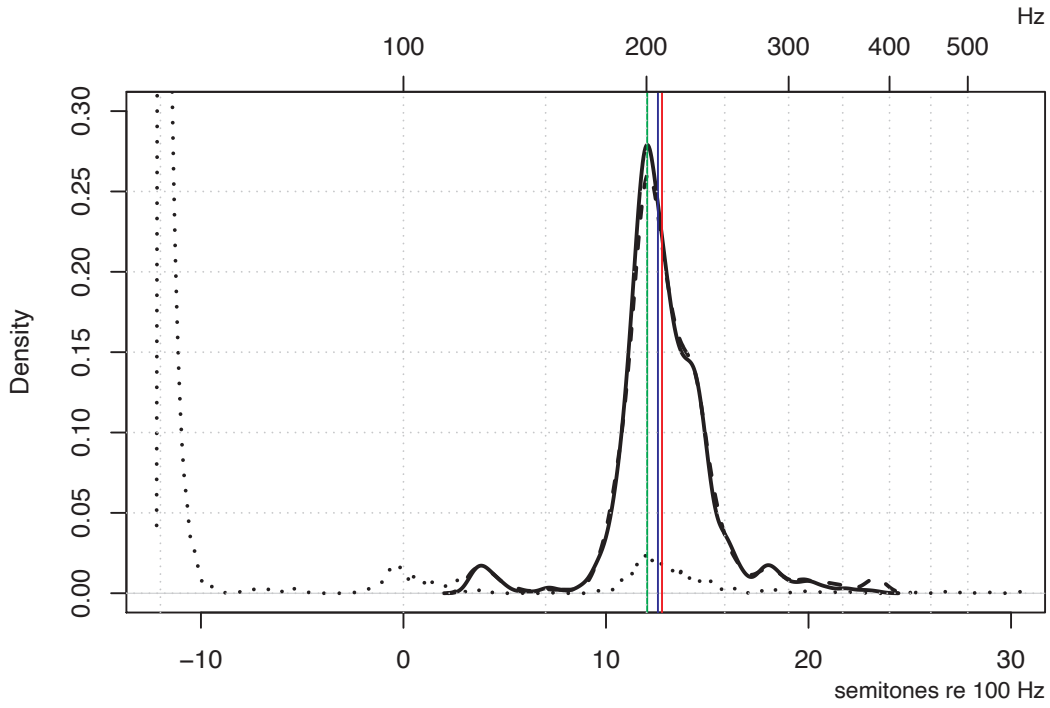

Figure 2: The pitch distribution of the speaker F23 (Corpus B), whose recording contained a constant, low, humming background noise at around 50 Hz. The effect of the noise is prominent in the raw overall pitch distribution (Dataset 1, dotted line), where the minimum frequency parameter was set at 50 Hz.

The number of pitch frames analyzed for each speaker is provided in Table 1. A summary of their individual pitch statistics in Dataset 3 is provided in Table 2. As a general observation, it is seen that the pitch mode for the maximal data in Dataset 3 is in most cases (for 33 speakers out of 40) located below the median, which in turn is usually below the mean pitch for each speaker. This confirms that a majority of the distributions are skewed to the right. Only seven speakers (F1, F21, M4, M5, M21, M22 and M26) are different in this respect. M4 has an almost symmetric distribution, and M5 is even slightly skewed to the left. Both of them creaked quite extensively. M21 and M22 exhibit bimodal pitch distributions, which may be due to the technical quality of the audio, perhaps overlapping speech. M26 has a relatively flat and irregular distribution.

Table 1: The number of pitch frames recorded for each of the 40 speakers in Dataset 3. The speakers of Corpus A are shown in the two leftmost columns, and speakers in Corpus B in the rightmost ones.

Speaker	N	Speaker	N	Speaker	N	Speaker	N
0F1	16335	0M1	13387	F21	17296	M21	10139
0F2	15669	0M2	13409	F22	04757	M22	07712
0F3	12640	0M3	12144	F23	02282	M23	04900
0F4	15455	0M4	21201	F24	07829	M24	02790
0F5	14563	0M5	15405	F25	08424	M25	02729
0F6	19197	0M6	14313	F26	08846	M26	05053
0F7	15506	0M7	19685	F27	12056	M27	02397
0F8	15615	0M8	16024	F28	17702	M28	08638
0F9	15778	0M9	21053				
F10	09725	M10	10216				
F11	15921	M11	19740				
F12	14737	M12	08217				

5.2 Establishing a reference pitch for comparing speakers

Figure 3 shows the pitch densities of all 40 speakers in Dataset 3. It is observed that male speakers tend to have lower pitch than females, which is hardly surprising. The overall mean pitch in Dataset 3 was 191.3 Hz (10.8 ST) for female speakers and 117.5 (2.2 ST) for males, with all speakers pooled. The corresponding standard deviations were 44.6 Hz (3.8 ST) for females and 33.0 Hz (4.5 ST) for males. The pitch distributions for the individual males form a cluster around 100 Hz or below, and most of the distributions for females are centered at about 150-200 Hz. However, there are also 6 male and 2 female speakers with more clearly overlapping distributions whose modes are located between 100 and 150 Hz. It is thus not uncommon for the two genders to exhibit similar pitch. Another important observation is that the shapes of these primary distributions exhibit at least roughly similar properties: usually one peak, generally similar width, and the distributions are more or less right-skewed.

In order to compare the way different speakers exploit their typical pitch range, it is possible to shift the pitch distributions over each other by referring the semitone-scaled pitch values to the speaker-specific modes. Using the semitone scale and the mode as the common anchor point enables us to compare the details of the individual distributions, while no information is lost about the perceptual distances of the pitch values. The result is shown in Figure 4.

Figure 5 shows a histogram of the mode-referred pitch values pooled for all 40 speakers in Dataset 3, supplemented with the corresponding probability density curve. The pooled mean of mode-referred pitch was 1.14 ST (s = 3.36 ST, median 0.61 ST). In the histogram of the pooled data, the probability of the bin with the highest probability (-0.5–0.5 ST) was 0.17 (17 %). The sum of the probabilities of the bins between -2.5 ST and 4.5 ST was approximately 0.77.

Table 2: Summary statistics of the primary pitch distributions for 40 speakers (Sp.) in Dataset 3.

Sp.	Mode		Median		Mean		Stdev	
	ST	Hz	ST	Hz	ST	Hz	ST	Hz
F1	10.29	165.84	10.23	180.59	10.55	186.35	2.73	31.85
F2	09.47	172.06	09.64	174.49	10.00	180.35	2.63	29.69
F3	09.68	174.38	10.60	184.48	11.27	194.85	3.03	37.30
F4	11.29	189.61	12.86	210.23	13.36	221.05	3.50	48.64
F5	12.03	198.85	12.94	211.21	13.21	216.45	2.30	30.51
F6	09.59	172.85	10.47	183.08	11.12	194.11	3.39	43.01
F7	09.97	177.08	10.64	184.91	11.27	194.34	2.72	35.74
F8	09.69	173.93	11.24	191.46	11.81	202.78	3.73	48.71
F9	07.89	156.78	08.88	167.01	9.48	175.96	3.14	35.92
F10	03.06	118.57	03.61	123.19	3.90	127.05	2.85	22.83
F11	06.05	137.95	06.31	144.02	6.65	148.53	2.60	23.83
F12	07.06	150.13	07.42	153.51	7.65	156.72	2.01	20.19
F21	13.79	192.77	13.09	212.94	13.24	217.74	2.75	37.12
F22	10.73	184.95	11.28	191.85	11.78	201.13	3.25	40.83
F23	12.03	200.33	12.56	206.63	12.77	211.2	2.49	30.66
F24	10.44	182.09	11.19	190.84	11.91	202.25	2.99	40.26
F25	09.95	176.42	11.75	197.12	12.56	213.03	4.19	56.82
F26	13.04	210.59	13.59	219.23	13.95	229.00	3.63	51.17
F27	08.79	161.00	09.56	173.74	10.15	183.16	3.25	38.96
F28	09.00	166.97	10.05	178.72	10.71	189.75	3.53	41.96
M1	-1.22	093.09	-0.07	099.61	0.60	106.08	3.60	27.57
M2	-0.50	096.65	00.57	103.34	1.15	108.60	3.05	21.10
M3	-0.55	096.37	00.73	104.33	1.65	112.67	3.66	26.69
M4	06.80	147.51	06.79	147.98	7.06	152.78	3.03	29.65
M5	-0.49	095.56	-0.85	095.19	-0.32	104.91	5.49	52.94
M6	01.76	110.22	02.76	117.27	3.31	123.75	3.49	28.50
M7	-1.45	090.50	00.08	100.44	0.77	107.24	3.75	27.06
M8	-6.01	069.79	-5.93	071.00	-5.54	074.23	3.49	17.14
M9	04.45	128.15	04.81	132.05	5.16	136.33	2.65	21.79
M10	-3.07	083.35	-1.40	092.23	-0.74	097.53	3.17	20.10
M11	-0.04	098.49	01.42	108.52	2.16	116.59	4.01	29.92
M12	04.64	130.14	05.23	135.23	5.64	140.03	2.49	22.31
M21	02.08	112.21	01.56	109.40	1.43	111.37	3.85	25.48
M22	04.49	112.49	04.34	128.51	4.44	131.28	3.04	23.74
M23	01.26	106.74	02.43	115.07	2.86	119.52	2.77	20.52
M24	-1.03	093.87	00.14	100.81	1.18	109.72	3.71	26.46
M25	-0.48	096.75	00.63	103.73	1.45	110.72	3.19	22.71
M26	03.56	110.54	02.59	116.17	2.56	118.09	3.32	23.20
M27	-1.21	092.78	-0.52	097.06	0.05	102.22	3.28	21.30
M28	01.94	111.28	02.81	117.63	3.21	122.50	3.18	23.55

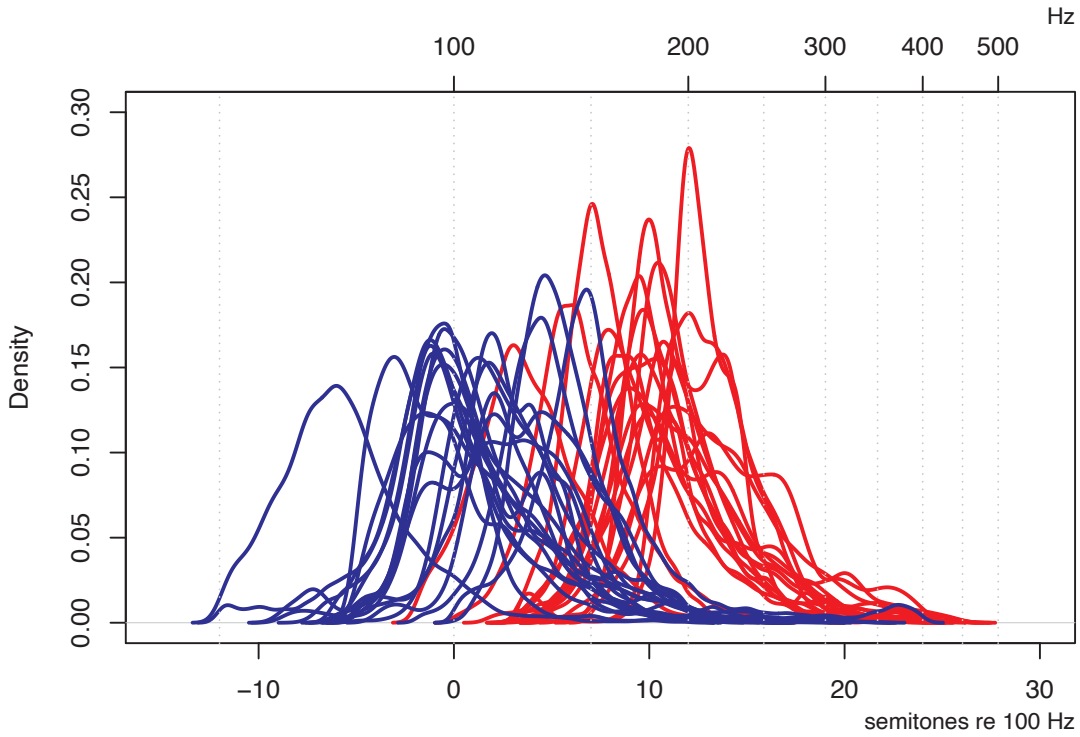

Figure 3: The overall pitch densities within annotated utterances of 20 male (blue lines) and 20 female (red) speakers.

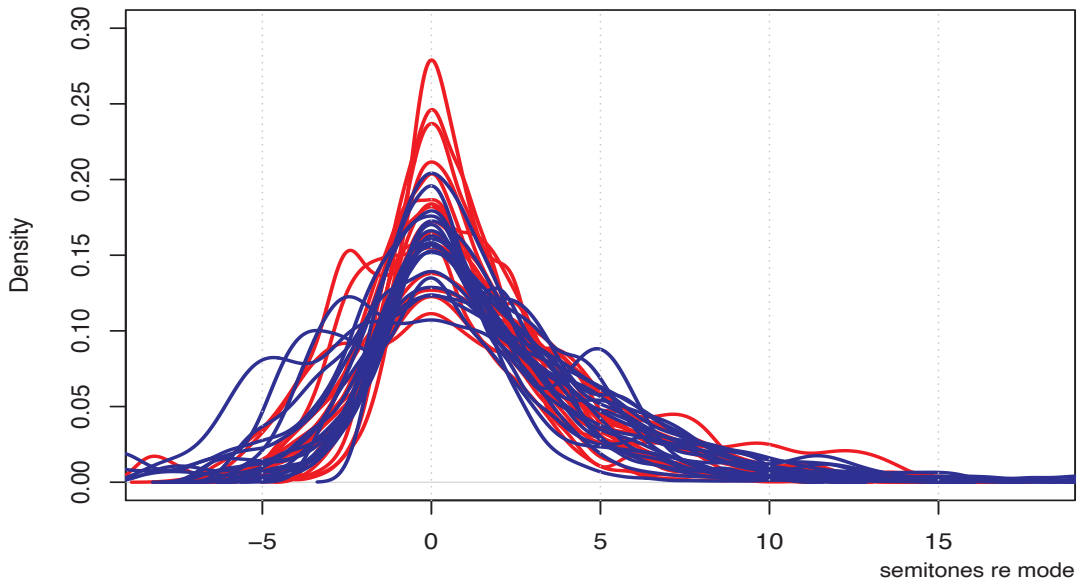

Figure 4: The mode-referred pitch distributions plotted as density curves for 40 speakers in Dataset 3. The zero pitch level refers to the speaker-specific mode. Male speakers are indicated with blue lines, females with red.

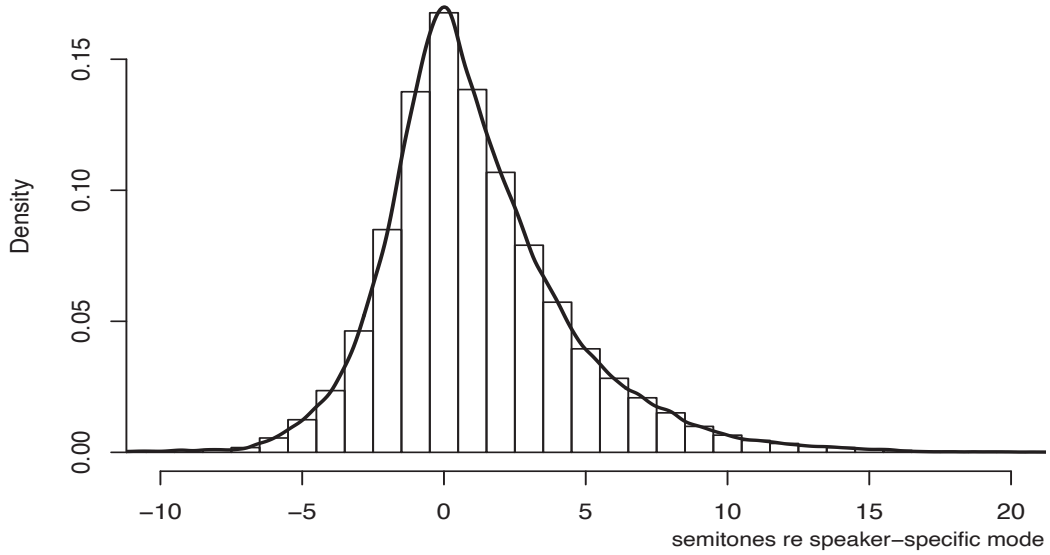

Figure 5: The distribution of mode-referred pitch values in voiced speech (N = 277,384) for all 40 speakers in Dataset 3. The zero pitch level refers to the speaker-specific mode. The bin width in the histogram is 1 semitone.

Thus, speakers would tend to exhibit pitch levels within such a span around their most typical pitch in about 77% of their voiced speech. 95% of all pitch values in Dataset 3 fall in the bins whose midpoints are located between -4 ST and 8 ST. Conversely, speakers would hit pitch levels outside this span in about 5% of their speech produced in the modal register. Since these probabilities are based on pooled data, they are to be taken as rough approximations. Speakers may differ to some extent, e.g., in the effective width of the primary pitch distribution.

5.3 Technical observations

Pitch analysis provides inconsistent results in cases where several speakers are captured in the same single-channel sound signal and two or more of them are speaking simultaneously. The analysis for the present study did not exclude the overlapped portions, since the amount of audible "crosstalk" in these dialogue corpora was considered relatively small and it only concerned a few speakers. However, such an exclusive feature could easily be implemented in the Praat script, when it is known which annotation tiers contain the utterance items that should not overlap.

The audio signal may sometimes contain background noise or electrical disturbances that can distort the pitch detection. For instance, in two of the dialogues in Corpus B, a humming noise was detected at the frequency of 50 Hz. This persistent noise is included in the analysis of Dataset 1 and thus creates an extra peak in the pitch distribution (see Figure 2 for an example). Since this kind of noise occurs within a low frequency range and usually does not overlap with speech frequencies, it might be possible to filter the noise out without significantly affecting the actual speech signal.

5.4 Bootstrapping

In order to estimate the minimum amount of speech that is required in order to obtain a reliable statistical description of the speaker's typical pitch range, we applied a bootstrapping procedure. In statistics, bootstrapping refers to any method – usually a statistic or a test – that uses random resampling of existing data. Bootstrapping can be used for calculating accuracy estimates of a (likewise estimated) statistic (Efron, 2003; Efron and Tibshirani, 1994). As such, it is used in finding sample sizes required for the convergence of a given statistical estimate that originates from an unknown distribution. In practice, random samples are drawn from a larger body of data. These samples are then analyzed as if they were regular samples from the studied phenomenon. For instance, it is possible to systematically increase sample size and repeat the random sampling a number of times for each sample size, and for each of these simulated samples to calculate the mean. This would provide a bootstrap estimate of the variation of the mean as a function of sample size and give us a way of estimating the sample size corresponding to a required level of accuracy.

For each of the 40 speakers, subsets of consecutive pitch values were randomly drawn from Dataset 3, beginning with the sample size of 50 pitch values (corresponding to 1 second of net speaking time) and increasing the sample size in steps of 50 values after each sampling round, either until the speaker had fewer samples than 1.5 times the sample size or until the maximum sample size of 10,000 pitch values was reached. For each sample size and for each speaker, up to five non-overlapping sequences of pitch values were drawn from the dataset, depending on whether a sufficient number of frames were available for the speaker in question. One single draw in the maximum sample size was possible for 16 speakers, who were represented with more than 15,000 pitch frames. The means of all the sampled portions from all 40 speakers are plotted in Figure 6, and the corresponding modes are shown in Figure 7. At sample sizes larger than 3000, fewer than five draws were possible for most speakers. However, the mean and mode have mostly converged before this point.

As shown in Figure 6, the standard deviation of the pitch means is about 2 ST in small sample sizes, but is reduced into less than 1 semitone after analyzing 650 pitch frames (only 12 seconds) or more. For many speakers, the pitch mode also converges quickly to a rather stable level and the overall standard deviation drops under 1 semitone after analyzing at least 34 seconds of net speaking time. For some speakers in Corpus B, the overall pitch distribution was bimodal, and the location of the primary mode is unstable, even after three minutes of net speaking time (e.g., speakers M21, M22, F21, F27). This phenomenon is visible in the mode-referred distributions that would overlap to a large extent apart from three female and two male speakers (see Figure 7). The bimodal distributions might be partly explained by the type of audio material. The recordings of the dialogues among M21 and F21, as well as F27 and F28, were noisy, the dialogues were recorded with only one microphone, and the speakers often overlapped in the signal. The reason for obtaining a bimodal pitch distribution in M22's recording was less clear, although background noise was present.

In some cases of Corpus B, the small amount of material available may explain why the distributions look unstable (cf. Table 1).

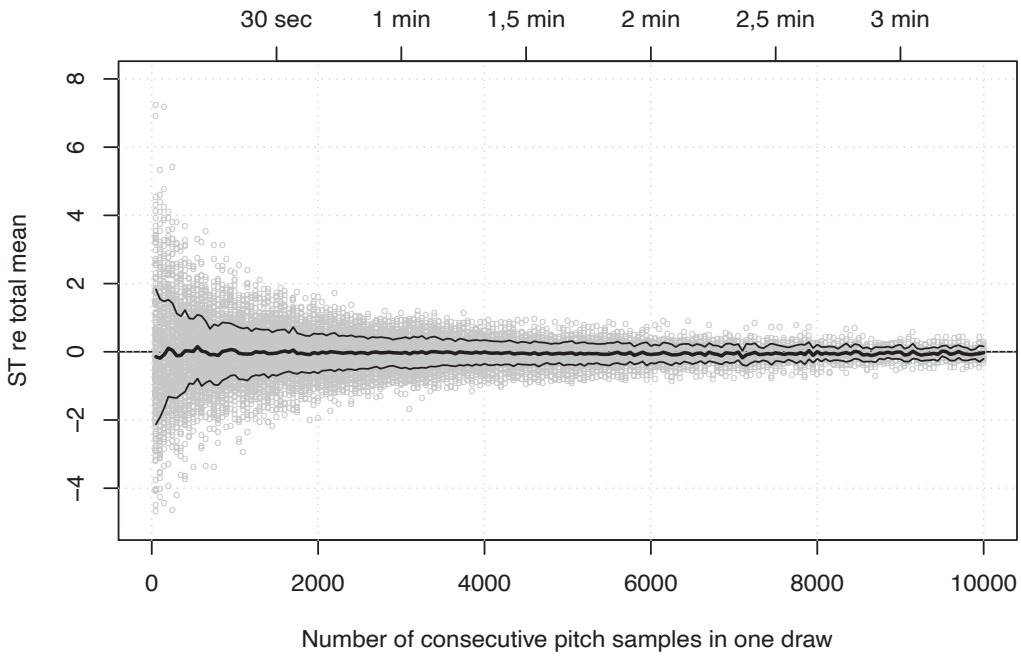

Figure 6: Bootstrapping the pitch mean. At most five sequences of 50 to 10000 consecutive pitch values were randomly drawn from each of the 40 speakers in Dataset 3. The **means** of all draws are plotted as grey circles relative to the corresponding speaker's total mean. The thick curve shows the local mean and the thin curves show the standard deviation for the means in each sample size. The values converge towards the speaker-specific mean in the complete dataset (zero level).

Figures 8 and 9 show the more detailed density curves in three exemplary conditions where each speaker is represented by one random sample of either 1000, 3000 or 6000 consecutive pitch points. In Figure 8, these pitch values are shown with respect to each speaker's overall pitch mean, and Figure 9 shows the corresponding mode-referred distributions. The mean of the pitch modes for the complete 1000-point samples was 0.12 ST (standard deviation 1.14 ST), 0.05 ST (s = 0.80 ST) for 3000 points and -0.11 ST (s = 0.43 ST) for 6000 points. The corresponding mean of the pitch means was 0.04 ST (s = 0.73 ST) for 1000, 0.04 ST (s = 0.44 ST) for 3000 and -0.08 ST (s = 0.25) for 6000 points.

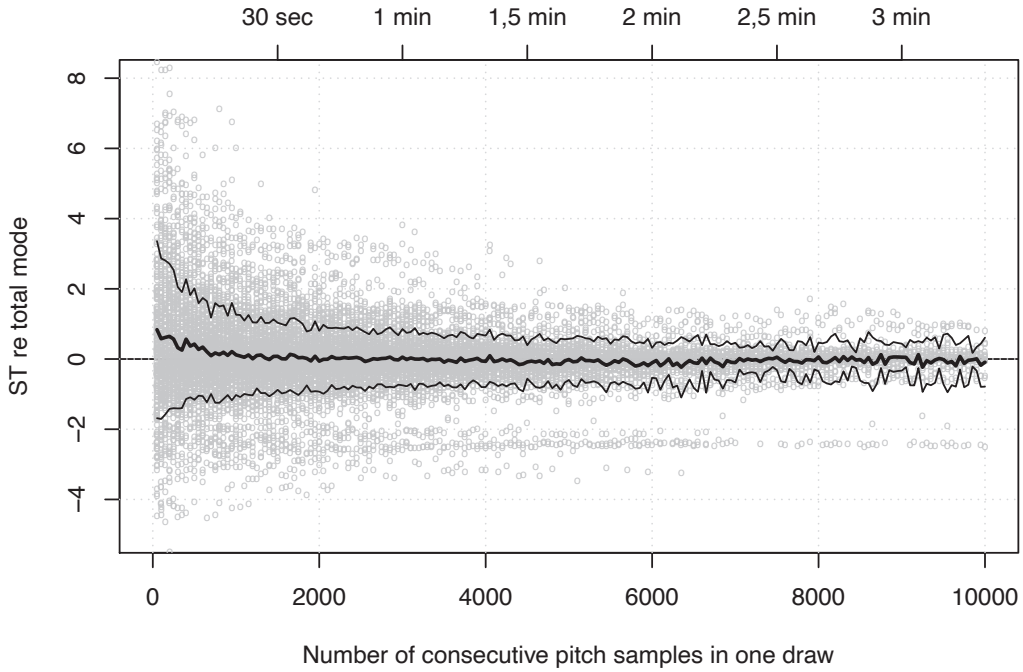

Figure 7: Bootstrapping the pitch mode. At most five sequences of 50 to 10000 consecutive pitch values were randomly drawn from each of the 40 speakers in Dataset 3. The **modes** of all draws are plotted as grey circles relative to the corresponding speaker's pitch mode in the complete dataset. The thick curve shows the local mean of the modes, whereas the thin curves show the standard deviation for each sample size. Four speakers (cf. the curves with "additional" peaks in the rightmost panel of Figure 9) exhibited bimodal pitch distributions, and their primary modes do not seem to fully converge even after 3 minutes of speech is included. These speakers contribute to the secondary "row" of data points below the overall mode.

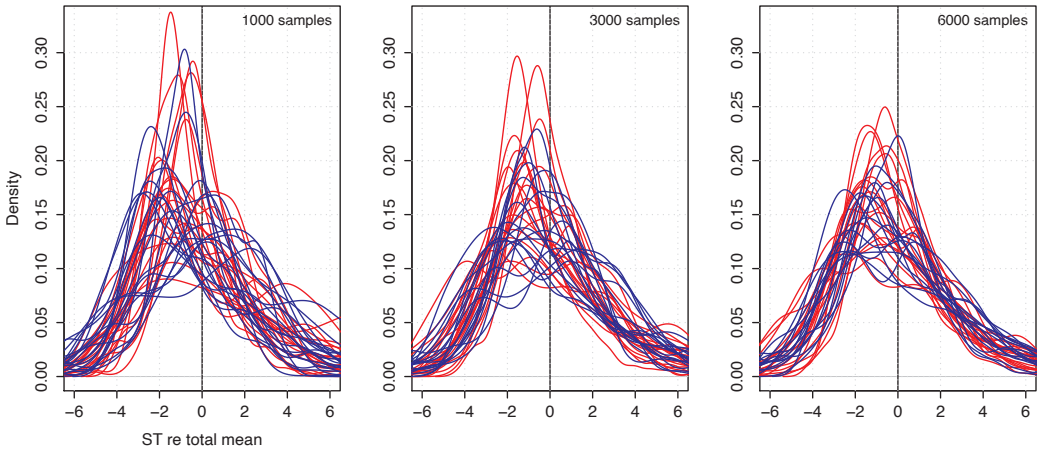

Figure 8: Distribution of a randomly selected subset of 1000, 3000 or 6000 consecutive pitch samples from 40 speakers. The pitch values are referred to the speaker-specific total **mean**, shown as the black vertical line in each plot.

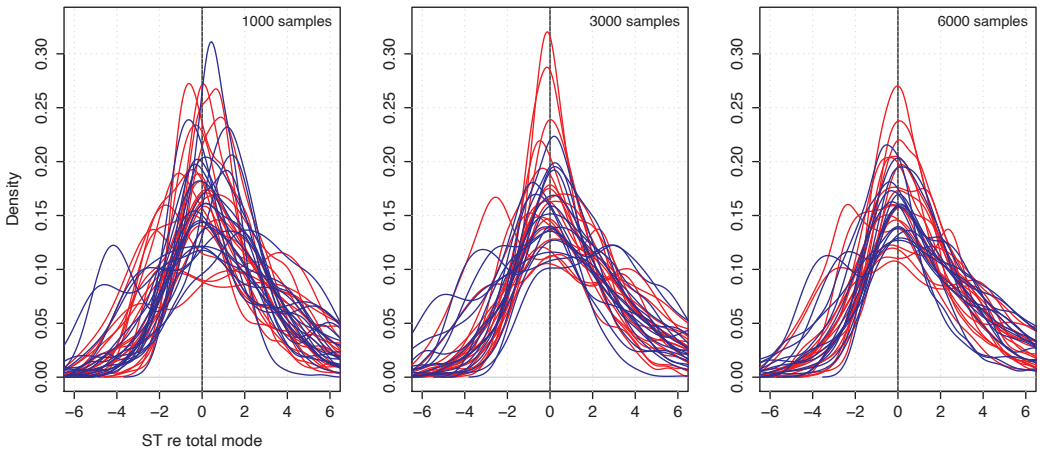

Figure 9: Distribution of a randomly selected subset of 1000, 3000 or 6000 consecutive pitch samples from 40 speakers. The pitch values are referred to the speaker-specific total **mode**, shown as the black vertical line in each plot. Male speakers are indicated with blue lines; females with red.

6 Conclusions

It was confirmed that in a sufficiently large dataset, a majority of the pitch values measured from each individual speaker tend to be distributed in a roughly similar fashion. This is likely to reflect the natural modes of vibration of the vocal folds and thus the pitch ranges of probable comfort vs. discomfort for the speaker. The primary distributions tend to be generally right-skewed. This observation is consistent with previous data (see, e.g., Traunmüller & Eriksson, 1995). The skewed distribution may be at least partly due to the fact that the length of the vocal folds sets a natural lower limit to glottal frequency, whereas humans can rather flexibly stretch their vocal folds in order to increase the pitch of their voices.

On the basis of these two corpora, it is typical for speakers to exhibit a primary pitch range that extends about 3–6 ST below and 6–12 ST above the pitch mode. Secondary "bulks" of data may be observed below and/or above the main range in the pitch distribution. In case these local modes occur at a distance of 12 ST (i.e., one octave) from the main pitch mode, it is to be suspected that they reflect a tendency of the speaker to use non-modal laryngeal settings (such as creaky voice or falsetto) and/or that the pitch analysis parameters have not been set in an optimal way for the speaker in question. For specific research purposes, it may be desirable to keep those results where the speaker's actual fundamental frequency has potentially been halved or doubled, since these may provide information about voice quality changes. In some cases, however, the additional modes may be due to other overlapping speakers or periodic background noise and need to be excluded. The present study paves the way for further research on the effects of various technical issues on pitch analysis, such as those of recording equipment, background noise, overlapping speech, voice quality differences, etc.

The minimum and maximum pitch do not provide a reliable summary of the speaker's preferred pitch range, since they are easily affected by non-modal voice

quality as well as by the selected analysis parameters. The standard deviation of the bootstrapped means and modes was reduced to less than 1 semitone after analyzing about 30 seconds or about 1500 pitch frames of net speaking time, given that the analysis parameters were set in an appropriate way. This may already be accurate enough for many research purposes. In case it is possible to determine the pitch mode of each particular speaker within a speech corpus, the mode is a good reference level for comparing the ways in which different speakers utilize their typical pitch ranges.

The tools for the analysis of pitch distributions may be applied in various domains, such as phonological models of intonation or clinical voice assessment. Given that some aspects in the pitch distributions may be highly speaker-dependent and relatively stable across different situations, the present tools may be applicable in the study of social identity (cf. Pierrehumbert et al., 2004; Munson, 2007; Cartei et al., 2014; Munson et al., 2015) and in the development of forensic speaker recognition (see Kinoshita et al., 2009). In terms of external factors that can affect speech, the tools for analyzing pitch distributions may be useful in studies of the effects of noise on speech production (cf. Hazan & Baker, 2011; Vainio et al., 2012) or for revealing whether speakers tend to accommodate their pitch levels to those of other speakers (cf., Gregory et al., 1993; 2001; Bosshardt et al., 1997; Babel and Bulatov, 2012; Garnier et al., 2013). Our findings will also be of interest in the analysis of the sequential unfolding of spoken social interaction, where the pitch range of the participants may systematically vary according to the position of a spoken turn within a larger sequence of turns (Stevanovic et al., submitted) and where speakers may be seeking to match each other's pitch levels according to sequential contingencies (Szczepek-Reed, 2010; Stevanovic & Lennes, submitted).

References

Babel M., & D. Bulatov D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech* 55, 231–248.

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer* [Computer program]. Version 5.4.17, retrieved 3 September 2015 from http://www.praat.org/.

Bosshardt H.-G., Sappok, C., Knipschild, M., & Hölscher, C. (1997). Spontaneous imitation of fundamental frequency and speech rate by nonstutterers and stutterers. *Journal of Psycholinguistic Research* 26, 425–448.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, 25, 1044-1098. doi:10.1080/01690965.2010.504378.

Cartei, V., Cowles, W., Banerjee, R. & Reby, D. (2014). Control of Voice Gender in Pre-Pubertal Children. *British Journal of Developmental Psychology* 32(1): 100–106.

Couper-Kuhlen, E. (1996). The prosody of repetition. On quoting and mimicry. In: Elizabeth Couper-Kuhlen & Margret Selting (eds.), *Prosody in Conversation*. Cambridge University Press, Cambridge, 366-405.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23(2): 283–292.

Efron, B. (2003). Second Thoughts on the Bootstrap. *Statistical Science* 18(2): 135–140.

Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

The FinINTAS Corpus of Spontaneous and Read-aloud Finnish Speech. URN http://urn.fi/urn:nbn:fi:lb-20140730194. [Speech corpus].

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, E., Emanuel A. Schegloff, S. A. Thompson (eds.), *Interaction and Grammar*. Cambridge: Cambridge University Press: 134-84

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural Correlates of Phonetic Convergence and Speech Imitation. *Frontiers in Psychology* 4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3769680/, accessed June 27, 2015.

Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication,* 40, 189–212.

Gregory, S. W., Green, B. E., Carrothers, R. M., & Dagan, K. A. (2001). Verifying the primacy of voice fundamental frequency in social status accommodation. *Language and Communication*, 21, 37–60.

Gregory, S. W., Webster, S., & Huang, G. (1993). Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language and Communication*, 13, 195–217.

Hazan, V. & Baker, R. (2011) Acoustic-Phonetic Characteristics of Speech Produced with Communicative Intent to Counter Adverse Listening Conditions. *The Journal of the Acoustical Society of America,* 130(4), 2139–2152.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America* 117(4): 2193-200.

Kaimaki, M. (2010). Tunes in free variation and sequentially determined pitch alignment: evidence from interactional organization. *Journal of Greek Linguistics*, 10/2: 213-250.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech Language and the Law,* 16(1), 91-111.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge Studies in Linguistics 79.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics,* 11, 373–382.

Lennes, M. (2007). On pitch and perceptual prominence in conversational Finnish speech. *Proceedings of the International Congress of Phonetic Sciences 2007, 6.-10.8.2007, Saarbrücken, Germany*, 1061-1064.

Lennes, M. (2016). pitch-distributions: Version 1.3. doi:10.5281/zenodo.45868

Lennes, M., Aalto, D., & Palo, P. (2008). Puheen perustaajuusjakaumat: Alustavia tuloksia. In: O'Dell, M. L., & Nieminen, T., eds., *Fonetiikan päivät 2008. Tampere Studies in Language, Translation and Culture, Series B,* 3, 147-155. Tampere: Tampere University Press.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America,* 102, 1864–1877.

Munson, B. (2007) The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation. *Language and Speech,* 50(1), 125–142.

Munson, B., Crocker, L., Pierrehumbert, J. B., Owen-Anderson, A., & Zucker K. J. (2015). Gender Typicality in Children's Speech: A Comparison of Boys with and without Gender Identity Disorder. *The Journal of the Acoustical Society of America,* 137(4), 1995–2003.

Ogden, Richard (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association,* 31(1), 139–152.

Persson, R. (2013). Intonation and sequential organization: Formulations in French talk-in-interaction. *Journal of Pragmatics,* 57, 19-38.

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The Influence of Sexual Orientation on Vowel Production (L). *The Journal of the Acoustical Society of America,* 116(4), 1905–1908.

R Core Team, 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.from http://www.R-project.org/.

Scherer, Klaus R., London, H., Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7(1): 31-44.

Stevanovic, M., Himberg, T., Niinisalo, M., Peräkylä, A., Sams, M. & Hari, R. (submitted).

Sequential approach to interpersonal synchrony: The case of joint decision-making.
Stevanovic, M. & Lennes, M. (submitted). Pitch matching – absolute or relative? On prosodic orientation across speaker changes.
Szczepek-Reed, B. (2004) Turn-final intonation revisited. In E. Couper-Kuhlen, & C. Ford (eds.): *Sound patterns in interaction: Cross-linguistic studies from conversation*. Amsterdam: John Benjamins. 97-117.
Szczepek-Reed, B. (2010). Prosody and alignment: a sequential perspective. *Cult Stud of Science Education,* 5, 859-867.
Titze, I. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America,* 85, 1699–1707.
Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. Unpublished manuscript. Retrieved 3 September 2015 from http://www2.ling.su.se/staff/hartmut/f0_mandf.pdf.
Vainio, M., Aalto, D., Suni, A., Arnhold, A., Raitio, T., Seijo, H., Järvikivi, J., & Alku, P. (2012). Effect of Noise Type and Level on Focus Related Fundamental Frequency Changes. In *Interspeech 2012 – 13th Annual Conference of the International Speech Communication Association,* Portland, Oregon, USA.
Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *TRASP'2013*, Aix-en-Provence, France. 7-10.
Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech,* 85(3), 315-337.