



VOICE CHARACTERISTICS OF FEMALE SPEECH AND THEIR REPRESENTATION IN COMPUTER SPEECH SYNTHESIS AND RECOGNITION

Dieter Huber

Department of Information Theory
Chalmers University of Technology
S-412 96 Göteborg
Sweden

ABSTRACT

This paper presents a comparative study of female *versus* male voice characteristics in read speech, both in a global (intonation) and local (laryngealization) perspective. Swedish newspaper texts comprising a total of 2610 running words were read by two female and two male speakers of Standard Swedish. The two women produced consistently more *intonation units* per text, with on the average higher F_0 onsets and offsets, shorter durations, steeper falls, a larger proportion of rising *versus* falling declination lines, and a markedly stronger tendency to time-align their intonation units with features of syntactic structure in the subsentence domain. Also, both female speakers made significantly more extensive and more varied use of *laryngealization* as a boundary marker than their male counterparts. The various voice parameters are described in quantitative terms with respect to their acoustical characteristics. Correlations and distributional properties are established in probabilistic terms for use in computer speech applications.

1. INTRODUCTION

Female speech has been shown to differ from male speech in several important respects including for instance higher average pitch (F_0), wider F_0 range (*key*), more symmetrical glottal waveshapes and consequently slightly steeper spectrum envelopes, a larger proportion of F_0 rises, more peripheral vowel articulations, and a consistently stronger tendency for women to avoid stigmatized speech variables and to produce more standard, grammatically and rhetorically "correct" utterances. Voice source and articulatory differences between female and male speakers are generally taken to be directly related to vocal tract size differences (i.e. larynx height and cavity, vocal fold length, vibrating mass, etc) between women and men. However, there are many more factors beyond mere anatomy and physiology which contribute to the characteristics of female speech, and it may not always be possible to determine with certainty the extent to which a specific feature derives from either physiological or linguistic differences, i.e. whether it represents an innate quality of female *versus* male speech, or rather the result of a learned behaviour reflecting social role, inter-language variation, style, or differences in the use of codes and dialects by women and men from the same speech community.

Quite obviously, an accurate representation of the voice source that captures not only linguistic variability and speech modes but also the inherent differences between female and male speech, is of paramount importance for all aspects of speech signal processing (analysis, synthesis, transmission, coding, enhancement, compression, etc) and computer speech applications (text-to-speech, speech recognition, speaker identification

and verification, etc). Earlier attempts to account for these differences in practical computer speech systems by employing simple scaling techniques (*viz.* fundamental frequency multiplied by a factor of 1.7, formants by a factor of 1.17) have not produced convincing results (see discussion in [1]). In order to arrive at an accurate representation of female voice characteristics in computer speech synthesis and recognition, clearly, both more parameters and larger ranges of potential variability have to be taken into consideration. In this paper, two voice source parameters are investigated in greater detail: intonation and laryngealization.

2. DATA

2.1 Texts

Three medium-sized Swedish newspaper texts (one narrative, one descriptive, one argumentative) comprising a total of 2610 running words, 176 graphic sentences and 65 paragraphs were selected as test material for this study. Average sentence length for the entire material was calculated to 14.8 words per sentence. The frequency distribution of sentence lengths (measured in number of words per sentence) is summarized in figure 1.

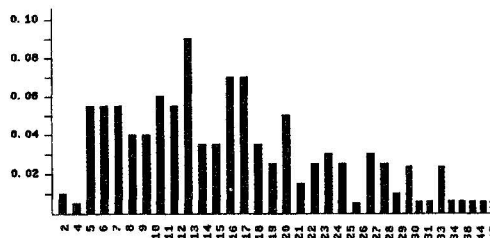


Figure 1 Sentence lengths in number of words

Type/token ratio for the accumulated material was calculated to 0.44, with a total vocabulary of 1152 lexical words. The full texts with complete listings of their respective constituent-based syntactic analyses, vocabulary lists and textual frequency ratings are contained in reference [2].

2.2 Speakers

Two female and two male adult native speakers of Standard Swedish participated in the reading tasks. None of the subjects reported any history of speech or hearing disorders. The speakers were chosen in an attempt to minimize possible dialectal differences and to represent a high standard of professionalism in oral reading skills.

Speakers LO (male) and LR (female) are professional radio journalists working with the Swedish broadcasting networks *Sveriges Radio*. Both of them are regularly engaged in news reading sessions as well as running their

own programs. Speaker BMH (female) is an airline employee working with passenger service and being used to making public announcements over loudspeaker as part of her daily working routines. Speaker SA (male) is a philologist and experienced public speaker who appears frequently in radio and television.

2.3 Methods

Registration of the speech samples was performed in an anechoic, sound-insulated recording studio, using a SONY PCM-F1 digital audio processor set to 16-bits quantization at a fixed sampling rate of 44.1 kHz. LPC-analysis and pitch extraction were performed using the SAP signal analysis package [3]. Pitch estimates were obtained at 16-msec intervals and calculated to the first decimal. Segmentation and broad classification of the resulting F_0 tracings into prosodically defined *intonation units* was performed by computer using the continuous speech segmentation algorithm described in [4]. This algorithm computes two global declination lines that approximate the trends in time of the peaks (topline) and valleys (baseline) of F_0 across the utterance. Computation is reiterated every time the *Pearson coefficient* drops below a preset level of acceptability. Segmentation is thus performed without prior knowledge of higher level linguistic information, with the termination of one unit being determined by the general resetting of the intonation contour wherever in the utterance it may occur.

The durations, *key* values, declination line slopes, and F_0 onsets (intercepts) and offsets (endpoints) of these intonation units, as well as their time-alignment with features of linguistic structure (constituent structure, sentence and paragraph boundaries, punctuation, etc) and physiological processing (pausing, breathing, etc) were established individually for each of the speakers participating in this study.

In addition, four different patterns of *laryngealization* were observed to occur consistently in the material. Following the classification scheme presented in [5] they were labeled as *glottalization*, *creak*, *creaky voice* and *diphthongic phonation*. The correlations of these patterns with different kinds of textually, syntactically, intonationally and physiologically induced junctures was established individually for each of the four speakers.

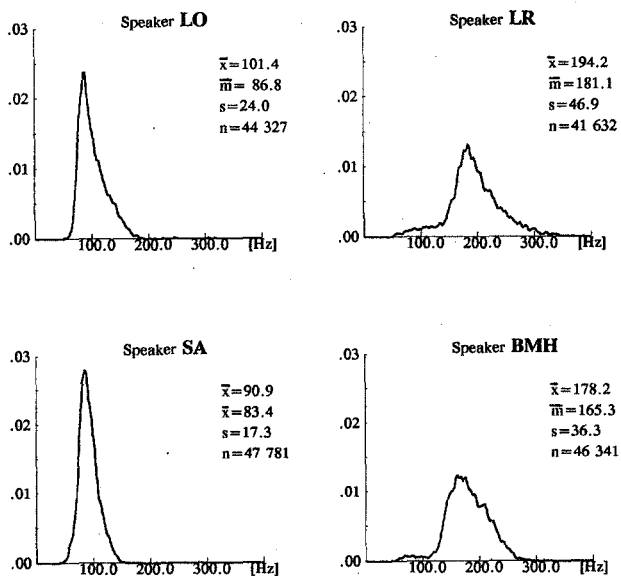


FIGURE 2 Fundamental frequency distributions

3. FUNDAMENTAL FREQUENCY DISTRIBUTION

Histograms of the voice fundamental frequency distributions (FFD) calculated separately for each of the four speakers for the accumulated text material are shown in figure 2. Also listed are the F_0 means (\bar{x}), modes (\bar{m}), standard deviations (s), and the number of 16-msec analysis windows containing voiced speech.

As can be seen in these FFDs, the average F_0 mean and F_0 variability range values are distinctly higher for the two women as compared with the two men. Individual values fall well within the margins proposed by other researchers (e.g.[6],[7]). It should be noted, however, that the modal values of F_0 , depicting the dominant, most frequent pitch values in the data as indicated by the peaks in the FFD histograms, are distinctly lower than the calculated means for each of the four speakers. This discrepancy is caused by the skewness of the F_0 frequency distribution. Clearly, the modal values (\bar{m}) are more representative of a speakers actual F_0 processing behaviour than the simple arithmetic means given by \bar{x} , and useful for practical computer speech applications.

In addition to the higher average F_0 and larger F_0 ranges in the female as compared with the male data, the FFD histograms depict one more apparently systematic difference between male and female voice, i.e. the existence of a clearly demarcated area of low and almost equally distributed F_0 values that lies distinctly below the "normal" range of pitch variability in the data for both female speakers, but is almost completely absent in the male histograms. This difference reflects the use of various patterns of laryngealization by our female and male speakers respectively, which will be discussed in more detail in section 5.

4. INTONATION

4.1 Number of intonation units

A total of 1664 intonation units has been established in the accumulated material for all four speakers. The distribution of these intonation units per speaker and text is summarized in table 1.

Table 1 Intonation units per speaker and text. Lines *n* state the number of occurrences, lines *r* give the ratio between *n* and the number of graphic sentences contained in the respective material.

		TEXT 1	TEXT 2	TEXT 3	TEXTS (all)
BMH	<i>n</i>	169	142	161	472
	<i>r</i>	3.38	2.95	2.06	2.68
LO	<i>n</i>	123	119	141	383
	<i>r</i>	2.46	2.47	1.81	2.18
LR	<i>n</i>	142	130	148	420
	<i>r</i>	2.84	2.70	1.89	2.39
SA	<i>n</i>	145	118	126	389
	<i>r</i>	2.90	2.45	1.61	2.21
Total (material)	<i>n</i>	579	509	576	1664
	<i>r</i>	2.89	2.65	1.84	2.36

As can be seen from these data, both female speakers BMH and LR produced more intonation units than their male counterparts LO and SA. This general tendency is stable in all dimensions, i.e. across the entire material (2.68/2.39 versus 2.18/2.21) and for each of the three texts separately (the latter with the single exception of SA reading TEXT 1 with the second highest score calculated for all four speakers, thus exceeding both LO and LR).

4.2 Correlations with linguistic structure

The correlations of the 1664 intonation units contained in the accumulated text material with sentences (S), clauses (C), nounphrase/subjects (SUB), verbphrases (VP), initial adverbials (ADV i), final adverbials (ADV f), parenthetical or parallel structures (PAR) and other (essentially non-grammatical) kinds of sentence structure (MIS) are listed separately for each of the four speakers in table 2.

Table 2 Correlation between intonation units and features of linguistic structure separately for each of the four speakers.

		BMH	LO	LR	SA	TOTAL
SENTENCE	<i>n</i>	67	74	63	95	299
	%	14.3	19.2	15.1	24.3	18.2
CLAUSE	<i>n</i>	184	157	173	148	662
	%	38.9	40.8	41.1	38.0	39.7
SUBJECT	<i>n</i>	37	12	19	15	83
	%	7.8	3.1	4.5	3.9	4.8
VERBPHRASE	<i>n</i>	28	13	17	18	76
	%	6.4	2.9	4.1	4.7	4.5
ADVERBIAL _i	<i>n</i>	16	3	11	5	35
	%	3.4	0.7	2.6	1.3	2.0
ADVERBIAL _f	<i>n</i>	39	37	33	32	141
	%	8.3	9.5	7.8	8.2	8.5
PARENTHETICAL	<i>n</i>	30	42	36	24	132
	%	6.4	11.0	8.6	6.2	8.0
MISCELLANEOUS	<i>n</i>	71	47	68	52	238
	%	15.0	12.3	16.2	13.4	14.3

Thus, compared with their male counterparts the two female speakers not only produced on the average more intonation units per text, but they also displayed a markedly stronger tendency to process single constituents in the subsentence domain in terms of separate intonation units. This tendency is most pronounced in the categories SUB and ADV_i, while no statistically significant differences between our male and female speakers could be established with respect to the - otherwise predominant - time-alignment between clauses and intonation units. The opposite trend appears to apply to full sentences, which are significantly more often aligned with one single, cohesive intonation unit in the male material than in the female one.

4.3 Declination line parameters

The declination line parameters onset (intercept), offset (endpoint), duration, and slope for the baselines and toplines respectively, were calculated separately for each of the 1664 intonation units. Not surprisingly, the two women produced consistently shorter intonation units than their male counterparts, with on the average higher onsets and offsets, steeper falls, and a significantly larger proportion of rising versus falling declination lines (compare reference [2] for detailed data). More unexpectedly, however, while the frequency distribution histograms of the topline onset and offset values for the male speakers are distinctly uni-modal and display for both LO and SA an essentially regular Gaussian distribution, the two female speakers not only display considerably larger ranges of variability with respect to their topline onset/offset values, but inside these ranges both BMH and LR have each two clearly defined and distinctly demarcated modal peaks. This "female" bi-

modality only occurs for the toplines (not the baselines) and is illustrated in figure 3 below for the topline onset (intercept) values.

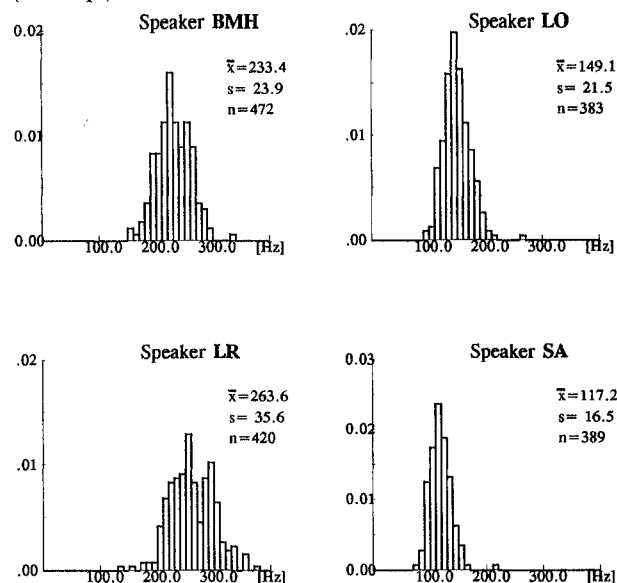


Figure 3 Frequency distribution histograms of topline-intercept values (in Hz) for the entire material. The slot intervals represent 10 Hz.

A complete listing of the exact figures, means, modes and standard deviations for all declination line parameters separately for each of the four speakers is contained in reference [2].

5. LARYNGEALIZATION

Four different patterns of laryngealization were observed to occur consistently at different kinds of textually, syntactically, intonationally or/and physiologically induced junctures. These patterns are exemplified below in figure 4 and will be denoted in the following as *glottalization*, *creaky voice*, *diplophonic phonation* and *creak*.

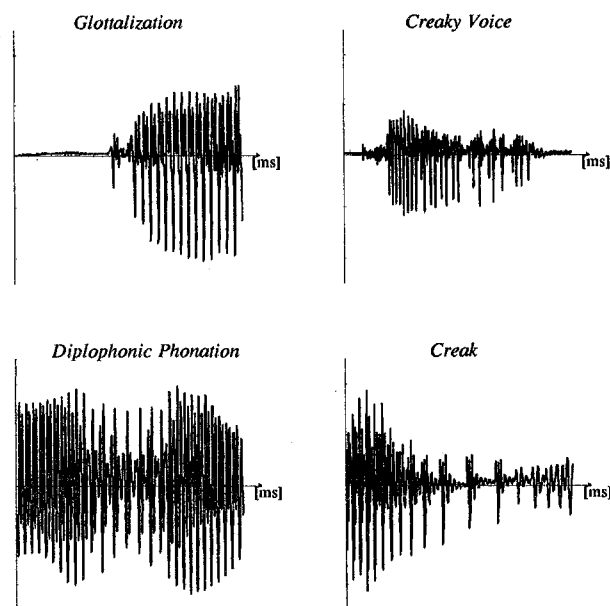


FIGURE 4 Patterns of laryngealization

All four speakers participating in this study produced instances of each of these four patterns of laryngealization at some time or another during the text reading sessions. The statistical frequency distributions of these patterns between the four speakers reveal, however, some interesting tendencies which are summarized in figure 5 and table 3 below.

Figure 5 Bar chart for speaker variability with respect to four different patterns of laryngealization. The number and percentage of occurrences for each pattern separately for each speaker are listed in table 3.

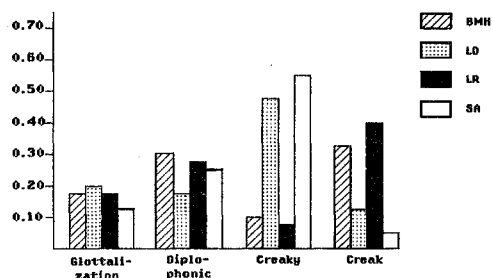


Table 3 Occurrences and distributions (per speaker) of different patterns of laryngealization in the entire text material (three texts). Lines *n* state the number of occurrences, lines % give the percentage figures calculated separately for each speaker.

		Glottalization	Diplophonia	Creaky-Voice	Creak	Total
BMH	<i>n</i>	76	134	89	142	441
	%	17.2	30.4	20.2	32.2	100.0
LO	<i>n</i>	41	33	93	24	191
	%	21.5	17.3	48.6	12.6	100.0
LR	<i>n</i>	69	109	58	157	393
	%	17.6	27.8	14.7	39.9	100.0
SA	<i>n</i>	33	61	138	17	249
	%	13.2	24.5	55.4	6.8	100.0

The following tendencies can be observed in these data:

1 - The two female speakers display considerably more incidences of laryngealization than their male counterparts. The total number of 1274 occurrences of laryngealization in the accumulated material is distributed as follows between the four speakers:

BMH	34.6%
LO	15.0%
LR	30.8%
SA	19.1%

2 - These average inter-speaker differences remain essentially stable in three of the four laryngeal patterns investigated in this study (i.e. *glottalization*, *diplophonic phonation* and *creak*), whereas the inverse relationship is found in the distributional data for *creaky voice* (i.e. the two male speakers display a significantly higher number of occurrences of *creaky voice* than their female counterparts).

3 - The distributional data for *creaky voice* and *creak* show some kind of reciprocity between the two male subjects on one side and the two women on the other, i.e. (our) speakers preferably use either creaky voice or creak at phonation offset (utterance final) positions, with (our) male speakers displaying a clear preference of creaky voice and (our) women an equally clear preference for

creak.

A closer study of the correlations of these four patterns with different kinds of boundaries revealed that creaky voice (for the men) and creak (for the women) occurred predominantly at pre-boundary, utterance-final, intonation unit offset positions, while glottalization occurred exclusively at post-boundary, utterance-initial (most often clause-initial), intonation unit onset locations. Diplophonic phonation, i.e. alternations between strong and weak glottal excitations occurred mostly at utterance-internal positions as a transition phenomenon between adjacent vowel sounds, thus serving as a kind of *laryngeal hiatus*.

SUMMARY AND CONCLUSIONS

Systematic differences between the two female speakers on one side and the two men on the other have been found with respect to practically all voice parameters investigated in this study. Compared with their male counterparts, the two women produced consistently more intonation units per text, with on the average higher onsets and offsets, shorter durations, steeper falls, a larger proportion of rising *versus* falling declination lines, and a markedly stronger tendency to time-align intonation units with features of syntactic structure in the subsentence domain. Further, both female speakers made significantly more extensive use of laryngealization as a boundary marker than their male counterparts. This tendency is most prominent for *diplophonic phonation* and *creak*, while the two men (in as much as they use laryngealization at all) apparently prefer *creaky voice*. The use of *glottalization* at post-boundary positions, on the other hand, does not appear to differ significantly between our male and female speakers.

Clearly, data based on two male and two female speakers of only one language variety, i.e. Standard Swedish, does not permit any cross-linguistically valid generalizations. Considerable larger amounts of data including more speakers, different speech styles, and other languages are necessary to verify the results presented in this study. Investigation of speech material produced by British English, French and Japanese female speakers is ongoing.

REFERENCES

- [1] D.H.Klatt, "Review of text-to-speech conversion for English", JASA 82(3), 1987
- [2] D.Huber, "Aspects of the communicative function of voice in text intonation", PhD dissertation, Göteborg 1988
- [3] P.Hedelin, "Manual of SAP-tasks", Technical Report 5, Chalmers University of Technology, Göteborg 1986
- [4] D.Huber, "A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units", *Proceedings ICASSP-89*, Glasgow 1989
- [5] D.Huber, "Laryngealization as a boundary cue in read speech", *Proceedings of the Second Swedish Phonetics Conference*, Lund 1988
- [6] G.Fant, "Acoustic theory of speech production", Mouton, The Hague 1960
- [7] V.W.Zue, "Acoustic theory of speech production", Supplementary notes for 6.343 Digital Speech Processing, M.I.T., Cambridge, Mass. 1985