Proceedings of the Institute of Acoustics

STABILITY OF AVERAGE LARYNGEAL FREQUENCY IN SPEECH

W. J. Barry (1), M. Goldsmith (2), A. J. Fourcin (1), H. Fuller (2).

(1) University College
(2) National Physical Laboratory

1. INTRODUCTION

Voice source characteristics are an important part of the "expressive" component in the speech communication chain, i.e. the component providing information about the identity and emotional state of the speaker [1]. The most basic, and easiest of these to characterise and measure is the frequency of vocal-fold vibration, which is considered both subjectively ([2], p.74, and [3]) and objectively ([4], p.82) important for the identification of a speaker. However, although there have been a number of studies devoted to the measure of average voice fundamental frequency in groups of 50 or more speakers (cf. [4], [5], [6] for survey information) and there is general recognition that a speaker's voice pitch varies with the situation, little has been done to clarify the general question of individual stability in that measure.

This paper presents results from an in-depth study[1] of the laryngeal patterns in speech of four speakers, and investigates the question of stability in a number of ways. Firstly it looks at average frequency data for varying stretches of speech from 10-15 seconds up to two minutes. There is evidence from the literature ([7], [8] [9], [10], [11]) that two minutes of continuous speech are sufficient to characterise a speaker, and this has been clinical practice in the UK for more than a decade. However, there are also indications that systematic longer term fluctuations exist [12]. Consequently, a series of longer (15-minute) recordings were made at different times of one day in order to check for a systematic shift in voice pitch during the day. Also, a number of different types of speech tasks were set in order to examine the stability of average voice frequency over tasks. Finally, the relationship between the "stability" of the two-minute period and the 15-minute period was examined.

2. SPEECH MATERIAL

Larynx and speech-signal recordings were made with four speakers (2F, 2M)[2] on two days under anechoic conditions using a two-channel Beta-Max recorder. Two of the speakers took part in 6 recording sessions of approximately 15 minutes each during the course of one of the two days. A "standard" extended reading task was given at 9.40am, 12.10pm, and 3.30pm (times ± 30 minutes), namely the reading of the "Environmental Passage" [13] followed by 13 minutes of continuous reading from a book of short stories. Three further tasks were set for comparison purposes: 1) Eight

[1]This study was carried out as part of Alvey Project MMI/132, Speech Technology Assessment. We are grateful for the advice on speaker selection criteria given by M. Keene, St Bartholomew's Hospital, and for the support in carrying out the recordings given by T. Sherwood, National Physical Laboratory and S. Nevard, University College. Results of the study are reported in [13], [14], [15].

[2]The four speakers were selected from a group of eighteen recorded in another set of recordings [13], [15] on the basis of their fluent reading ability.

consecutive readings of the "Environmental Passage" were recorded to provide a comparison with the unprepared reading from a book by keeping the textual structure constant and avoiding fluctuations in the interest and excitement of the contents. 2) Each speaker spoke freely for 15 minutes on a subject of his/her choice; this gives a longer term comparison between read and freely spoken monologue. 3) Two speakers at a time were recorded in dialogue; this adds a further dimension for comparison both with reading and with free monologue. Table 1 summarises the timetable for the four speakers.

Table 1. Speech tasks and times for the long-term larynx study

TASK	Day 1		Day 2	
	TJ	FC	CP	AH
1. Passage + book	9.10	9.40	9.33	10.12
2. Repeat passage	10.02	10.21	10.45	11.07
3. Monologue	11.01	11.22	11.27	11.47
4. Passage + book	11.46	12.05	12.06	12.25
5. Dialogue	12.30		12.46	
6. Passage + book	15.10	16.07	15.00	15.20

3. ANALYSIS PROCEDURES

Analysis of the larynx signal (Lx) was carried out in the following manner [16]: The duration (Tx) of each laryngeal cycle was determined, using a Voiscope ^{R}, which sampled the intervals between successive Lx waveform peaks at 500 kHz. The Tx values were then converted to Hz (to give Fx, the period-by-period frequency).

Frequency distributions. The frequency distributions (Dx) were calculated using 128 logarithmically equal bins on a scale from 30 to 1000Hz. Second-order distributions were used for all quantitative statements and statistical calculations in this study. They were derived by including an Lx cycles only if its duration is within 10% of that of the preceding cycle. This has the effect of eliminating the contributions of laryngeal irregularities from the distribution while retaining a maximum number of data points. In addition, the distributions used for the basic statistics are time weighted; that is, the proportion of time spent at a particular frequency is calculated rather than the number of laryngeal cycles of a particular frequency. This is done for two reasons. Firstly, our auditory impression of pitch movements is based on the frequency changes in time, and secondly, most voice-frequency studies have obtained their data from fixed *frame*-based rather than from fundamental- or laryngeal-cycle *period*-based analysis

Processing of the 15-minute sessions was done on sections of approximately 2 minutes, and overall distributions for the whole session was obtained by cumulation of the shorter sections. For the question of short-term stability, the recording of the (approximately 2-minute) reading-passage from the second standard session (task 4) was processed a second time in short sections of approximately 10-15 seconds.

The criteria for determining the length of the sections were a) to have roughly equal lengths of section covering the whole passage, and b) to have roughly equal numbers of laryngeal periods for the two men and the two women in each section in order to compare the progress towards stability and the fluctuation from section to section. There was some variation in the number of sections processed per speaker as a result of considerable differences in the elapsed time for the passage. For one speaker, TJ,

who has a extremely low laryngeal frequency, the "short" sections were considerably longer than 10 seconds in order to process a sufficient number of laryngeal cycles to make the measures meaningful in comparison.

4. RESULTS

Two aspects of larynx frequency stability are addressed: Firstly, the minimum duration of recording necessary to achieve a stable measure, secondly, the extent to which a "stable" measure is still liable to fluctuate in the longer term as a result of genuine baseline changes.

The question of achieving a stable measure which can be considered to characterise a person's voice is not a purely statistical one. If consecutive laryngeal cycles were independent (which they clearly cannot be), and if they were normally distributed, stability could be assumed after about 800 cycles ([17], p. 345). Fluctuation of laryngeal measures is dependent on the prosodic structure of the speech being analysed. Given the limited number of intonational patterns that exist in a language, there is a statistical element in the amount of speech that is necessary to achieve a representative sample of these patterns. However, the units of the pattern are not defined in physical terms, but are dependent on the intonational model adopted. The question is therefore approached empirically.

4.1 Short-term stability

The mean Fx values for short sections of a longer text against the mean for that text, and the cumulative mean of the short sections show a) the degree to which the short sections vary around the mean, and b) the progress of convergence on the overall mean with increasing duration. Figure 1a-d shows the average Fx values for the short sections of the reading passage.

Individual sections differ up to 23 Hz in the women speakers (FC vs. just over 12 Hz for CP) and up to 18 Hz in the men (TJ vs. over 12 Hz for AH). This represents a 3.3 semi-tone (ST) shift for TJ, 1.8 ST for both FC and AH, and 0.9 ST for CP. The cumulated average Fx values show that the 7th section (approx 80 seconds of speech) is the latest point (see speaker FC) at which the final frequency value ± 1% is reached. Speaker TJ reaches this threshold by the 6th section, AH after the 5th, and CP after 2 sections. This result compares quite well with those reported in the literature [10], [11] though the results for speaker FC and TJ suggest caution in restricting analysis to anything less than two minutes of recording.

The effects on the overall mean of extreme short-term fluctuations in mean frequency, such as those found for TJ, decrease with the duration of the recording, and a 2-minute passage therefore appears to be a safe basis from which to calculate personal voice-frequency characteristics. If the most extreme divergence from the overall mean had occurred during the last section of each speaker's reading, it would only have caused a shift of more than 1% in the mean for TJ, whose very low voice frequency, coupled with his low number of laryngeal cycles, meant that each section is weighted more than the other speakers' sections.

4.2 Medium-term stability

A similar approach to that taken for short-term stability can be applied to a speaker's medium-term voice-frequency behaviour. Figure 2a-d plots the mean values for approximately 2-minute sections of the first 15-minute standard reading (reading passage followed by 13 minutes reading from a book) against the overall mean for the

AVERAGE LARYNGEAL FREQUENCY

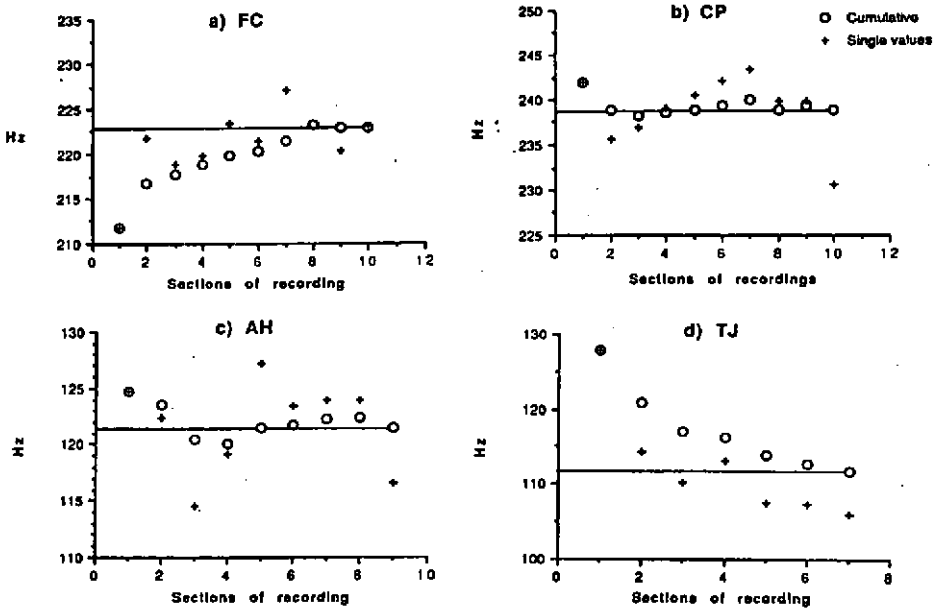

Figure 1a-d Mean Fx values for individual short sections, and cumulated mean values set in relation to overall mean for the reading passage from task 4

task. Here, as in the short-term comparison of individual sections with the overall mean, there are considerable differences between the mean values for the 2-minute sections. Table 2 gives the range in semi-tones between the lowest and highest mean frequency for a two minute stretch during each of the tasks.

Table 2 Difference in semi-tones between the highest and lowest mean frequency for a two-minute stretch during each 15-minute task.

TASK	FC	CP	AH	TJ
1. Passage + book	1.26	0.45	0.53	1.15
2. Repeat passage	1.36	0.19	0.72	0.91
3. Monologue	0.51	1.05	1.93	0.60
4. Passage + book	0.75	0.81	1.35	2.12
5. Dialogue	1.56	0.95	0.66	0.19
6. Passage + book	0.91	0.82	0.87	0.87

The largest pitch shifts appear to be randomly distributed over tasks and speakers. The comparison of the repeated-passage reading (task 2) with the other conditions, where the effect of e.g. content-induced pitch change would be maximised, does not support the assumption that shifts in mean frequency are a function of text-type alone. The difference for Task 2 is very low for speaker CP, but not for the other speakers, and in fact has the highest value of all tasks for FC. Figure 3a-d shows the mean values for the individual 2-minute stretches in relation to the overall mean for the repeated passage task (Task 2). The pattern is in general more regular than for the book readings, but in no uniform way across speakers.

AVERAGE LARYNGEAL FREQUENCY

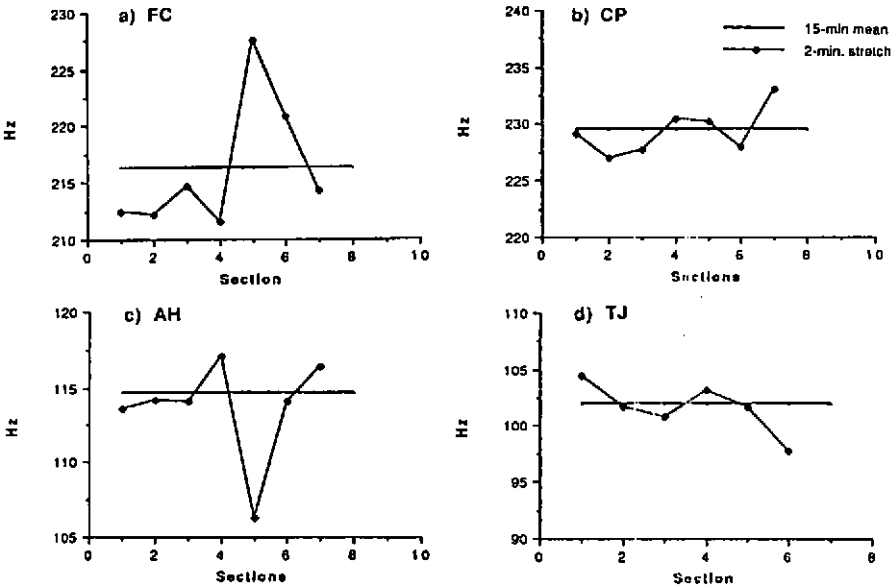

Figure 2a-d Mean Fx values for approximately 2-minute sections of reading (task 1) in relation to overall 15-minute mean

The differences between the speakers in the repeated-passage task, and the random distribution of large and small pitch differences for 2-minute stretches across the different tasks, indicate that there is, strictly speaking, no such thing as a generally valid personal voice-frequency value (compare [3]). Each situation has its specific and individual effect on a speaker's voice frequency. It is feasible that speaker FC experienced a growing tension during the repeated reading task in contrast to the other three speakers, in particular CP, whose mean voice excitation frequency remained very stable throughout the task.

4.3 Long-term stability

The variation in voice-frequency values within an extended speech production task of 15 minutes raises the question of comparability in the longer term. It is interesting in this context to compare the values found for the standard passage in this study with those for the same speakers reading the same passage during the Normative Study recordings some weeks earlier [13] Table 3 lists the values for all the *single* readings of the passage (i.e. for the Normative recording, and for first part of tasks 1, 4 and

Table 3 Mean Fx values (Hz) for the standard reading passage at the beginning of Tasks 1, 4 and 6.

	FC	CP	AH	TJ
Normative recording	208.3	224.7	129.9	101.5
Task 1 passage	212.5	229.1	113.6	104.4
Task 4 passage	222.9	238.8	121.5	109.4
Task 6 passage	228.2	233.6	123.1	109.7

AVERAGE LARYNGEAL FREQUENCY

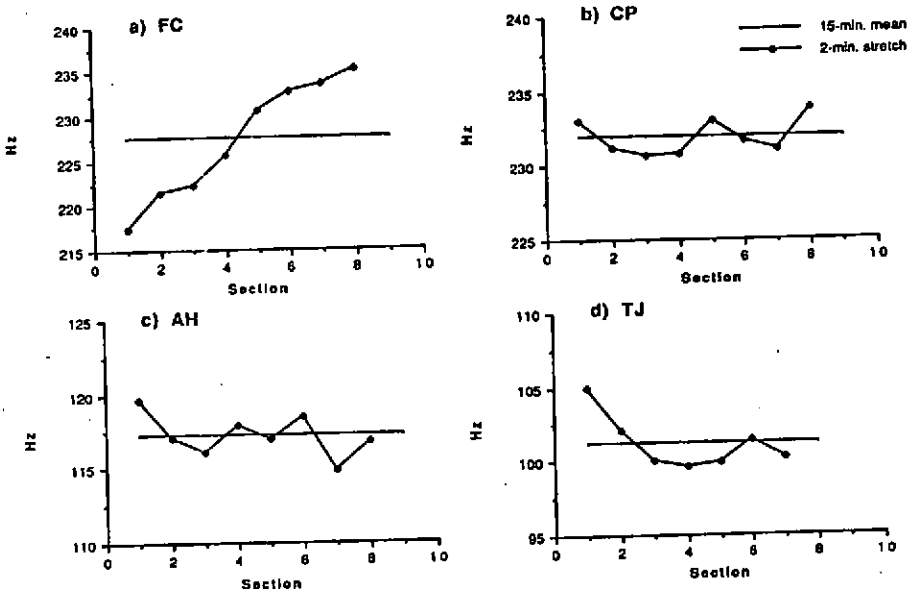

Figure 3a-d Mean Fx values for approximately 2-minute sections of repeated passage reading (task 2) in relation to overall mean

6 of this study). The overall picture of variation from situation to situation is confirmed by this overview. For each speaker, values for the same text vary from one recording to the next as much as the different 2-minute stretches varied within the 15-minute readings from a book. However, in table 3 there is a certain regularity in the values from task 1 to task 6. For all four speakers there is an increase from task 1 to task 4. This is followed by a fall-off in task 6 for CP while TJ maintains approximately the same mean excitation frequency and FC, and to a lesser extent AH, show another increase. Given the timing of the tasks (morning, mid-day, afternoon) this progression is similar to that reported by Garrett and Healey [12] who found an increase in fundamental frequency for all speakers during the first part of the day (9am - noon). In partial contrast to our results, however, they also found an increase for their group of male speakers during the afternoon with a fall-off restricted to female speakers.

These shifts are related to average frequency over longer stretches in figure 4a-d. The mean frequency values for the course of the approx. 2-minute stretches over the complete 45 minutes of reading are plotted against both the overall 45-minute mean and the 15-minute mean values. The periods of 15 minutes associated with the times of day are joined by a horizontal line, allowing visual examination of the frequency fluctuations within, and shifts across those periods. Three of the four speakers show a marked shift from morning to mid-day and fall back from mid-day to afternoon. This corresponds to part of the pattern found for the reading passage alone (compare table 3), though there was no clear fall-off in the afternoon. The one speaker (TJ) who has no marked shift in overall mean pitch from morning to mid-day *does* have a signifi-

cant shift from mid-day to afternoon (Mann-Whitney, U' = 6, p = 0.05). It must be stressed, however, that differences between individual stretches within the 15-minute periods can exceed the shifts between the periods (see speakers TJ and FC, though in neither case does this affect the significance of the shift. FC: U' = 4 < 10. p < 0.05)).

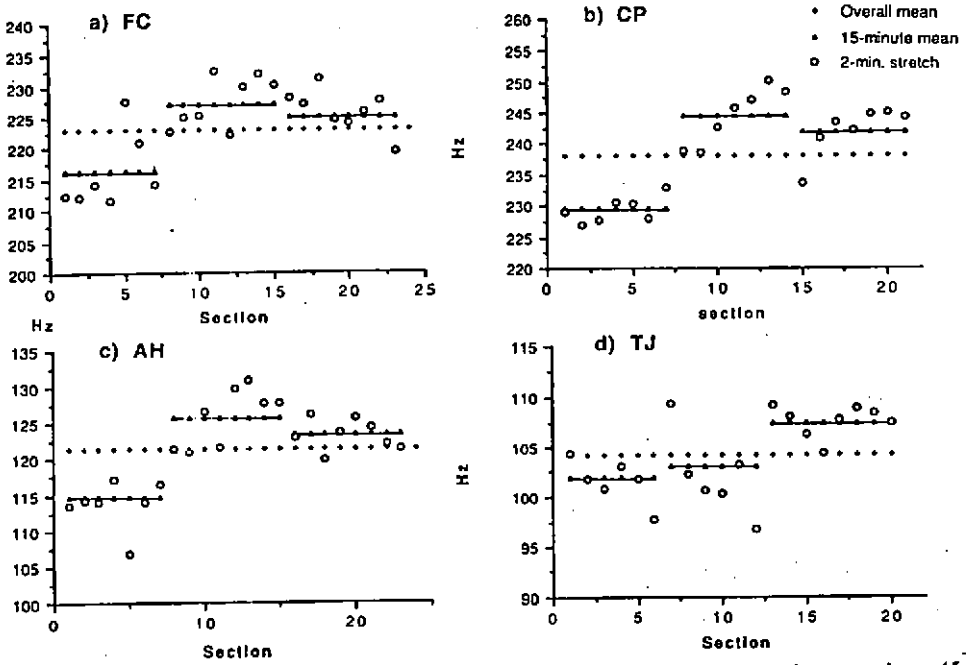

Figure 4a-d Mean Fx values for the approx. 2-minute sections over the complete 45 minutes of reading plotted against both the overall 45-minute mean and the 15-minute mean values (tasks 1, 4, 6)

4.4 Task dependence

Some indication of differences in voice frequency as a function of task has already been given with the comparison of task 2 (repeated reading of passage) with the reading from a book. Two other speech tasks were given to the subjects which extend the task range, namely free monologue and dialogue. Table 4 gives the overall mean values for monologue, dialogue, repeated reading, and free reading (compare the latter with the table 3 values for the separate free reading sessions).

Table 4. Overall mean values (Hz) for approx. 15 min. monologue, 15 min. dialogue, 15 min. repeated-passage reading, and 3 x 15 min. free reading.

TASK	FC	CP	AH	TJ
3. Monologue	191.2	237.2	113.9	89.2
5. Dialogue	209.3	222.2	104.1	99.9
2. Repeat passage	227.5	232.0	117.3	101.2
1,4,6. Pass. + book	223.3	238.8	121.4	104.2

AVERAGE LARYNGEAL FREQUENCY

There is a strong tendency for voice frequency to be lower for spontaneous than for read speech. All four dialogue values and three of the four monologue values (exception CP) are lower than for any of the reading conditions. This finding parallels the results from the STA Normative Study [13] and those from the literature discussed in it. It is interesting to note that in that study also, CP's voice frequency for the monologue was higher than for the reading tasks (238Hz and 240Hz for the two reading passages, 243Hz for the monologue).

5. SUMMARY AND CONCLUSIONS

The results from this investigation of voice-frequency stability indicate very clearly that while there is some justification for using mean frequency as a measure to characterise a person's speech if the sample is 2 minutes or more in duration, fluctuations from one sample to another forbid the use of such a measure as an absolute personal characteristic. It can only serve as a characterisation of the sample in question. Across speech tasks, and even between different 2-minute samples of the same extended task, mean frequency was shown to vary by as much as 15%. Within speakers, some regularity was found in mean-frequency change during the course of the day, and between spontaneous and read speech. This supports previous findings in the literature, but individual variation in the patterns found indicate that these trends also need to be treated with caution.

In the light of these results, the conclusion is unavoidable that reliance on mean voice frequency as an indicator of personal identity is inadvisable. A possible alternative measure is modal frequency, which though not insensitive to task variation is considerably more stable [14].

6. REFERENCES

[1] K. Bühler, Sprachtheorie. Jena, (1934)
[2] T. Broeders and T. Rietveld, Segmental markings as a cue in auditory voice identification of telephone speech. Proceedings of Eurospeech 89. European Conference on Speech Communication and Technology, Paris, September 1989. pp. 71-4. Edinburgh: CEP, (1989)
[3] E. Abberton and A. J. Fourcin, Intonation and Speaker Identification. Language and Speech 21, 305-318 (1978)
[4] H. J. Künzel, Sprechererkennung. Grundzüge forensicher Sprachverarbeitung. Heidelberg, (1987)
[5] A. E. Aronson, Clinical Voice Disorders. An Interdisciplinary Approach. New York, (1980)
[6] R. J. Baken, *Clinical Measurement of Speech and Voice*. Boston: College Hill Press, (1987)
[7] M. Steffan-Batog, W. Jassem and Gruszka-Koscielak, Statistical distribution of short-term F0 values as a personal voice characteristic. In: Jassem. W. (ed.) Speech Analysis and Synthesis 2, 195-206, Polish Academy of Science, (1970)
[8] N. Green, Automatic speaker recognition using pitch measurements in conversational speech. JSRU Report 1000, Joint Speech Research Unit, Ruislip, Middlesex, (1972)
[9] K. O. Mead, Identification of speakers from fundamental frequency contours in conversational speech. JSRU Report 1002, Joint Speech Research Unit, Ruislip, Middlesex, (1974)
[10] J. D. Markel and S. B. Davis, Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. IEEE Transactions, Acoustics, Speech and Signal Processing, ASSP-25, 330-337 (1979)

[11] S. Hiller, J. Laver and J. MacKenzie, Durational aspects of long-term measurements of fundamental frequency perturbations in connected speech. Work in Progress 17, 59-76, Dept. of Linguistics, Univ. of Edinburgh, (1984)

[12] K. L. Garrett and E. C. Healey, An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of the day. J. Acoust. Soc. Amer. 82 (1), 58-62, (1987)

[13] Barry, W. J., Goldsmith, M., Fourcin, A. J., and Fuller, H. (1990a): Larynx Analyses of Normative Reference Data. Project Report, Alvey Project MMI 132. London: University College.

[14] Barry, W. J., Goldsmith, M., Fourcin, A. J., and Fuller, H. (1990b): Stability of Laryngeal Measures in Speech. Project Report, Alvey Project MMI 132, London: University College.

[15] H. C. Fuller, A. J. Fourcin, M. J. Goldsmith and M. Keene, A database of normative speech recordings. Proceedings of Institute of Acoustics 1990 Autumn Conference, Speech and Hearing.

[16] A. J. Fourcin, Laryngographic assessment of phonatory function. In: C.L. Ludlow (ed.) Conference on the Assessment of Vocal Pathology, Maryland: ASHA Reports 11, (1981)

[17] F. E. Croxton and D. J. Cowden, *Applied General Statistics*. London, New York, Prentice Hall, (1951)