RESULTADOS DE INVESTIGAÇÃO PUBLICADOS NAS

ACTAS DO CONGRESSO "EUROSPEECH CONGRES "

EM SETEMBRO DE1989 EM PARIS

QUANTITATIVE STUDY OF THE EFFECTS OF SETTINGS CHANGES ON THE LTAS

Bernard HARMEGNIES (+), John ESLING (*), Véronique DELPLANCQ (+)

(+) Université de Mons, Mons, Belgium
(*) University of Victoria, Victoria (BC), Canada

ABSTRACT

Several experiments have suggested that deliberate changes of the speaker's voice quality result in modifying the LTAS of his speech productions. The few studies in this field have mainly dealt with speaker recognition tasks under various conditions, where the speakers were simply asked to modify their voices. However, Nolan (1983) had 2 trained subjects utter 30 times an invariant text, using a given 'setting' in each utterance. He computed a LTAS for each utterance and mainly performed a qualitative study of the spectral traces. The purpose of our experiment is to combine Nolan's approach, allowing to control the introduced vocal variability, with a quantitative technique allowing to measure the resulting spectral variability. One trained subject has produced a French balanced text with 31 different qualities. Each utterance has been analysed by means of a 2033 Bruel Kjaer FFT analyser. The so obtained LTAS were compared by means of the SDDD dissimilarity index. The results show what groups of settings deliver similar spectral results.

1. INTRODUCTION

Many experiments have shown that the long term average spectrum (LTAS) is a good acoustical cue to the speaker's individuality [1-4]. In other words, one can consider that the intra speaker variability is small in respect to the inter speaker variability. However, most of the work carried out in this field has dealt with speakers involved in quite artificial communication situations, i.e., most often, reading a text in a laboratory environment. Moreover, the speakers were usually asked to speak as naturaly as possible, so that the involved situations led them to minimize the variability of their voices. Therefore, it would be more convenient to state that it is the intra speaker variability of the natural, ordinary voice, which is small in respect to the inter speaker one.

Nevertheless, it would be a truism to recall that the human voice is very plastic : any speaker can very easily modify his or her own voice to a great extent. Few research dealing with voice quality have focussed on this topic. Yet, from a theoretical point of view, it could be very interesting to maximise the voice variability, in order to emphazise phenomena which could be more easily observed and therefore, could help to build a better understanding of the LTAS structures.

Hollien and his collaborators [5] have carried out experiments where subjects were asked to disguise their voices. They report important degradations of the recognition scores, suggesting major changes in the LTAS structures. Nevertheless, the LTAS are neither shown nor described, so that it is impossible to know what kind of changes occur. Moreover, as subjects were free to choose their disguises, it is impossible to control the kind of vocal changes actually performed. From this experiment, one can only infer that deliberate changes of voice quality may result in modifying the LTAS, but it is impossible to determine in what way the modifications occur.

From this point of view, the experiment by Nolan [6] seems more attractive : he had two trained subject utter several times the same text, each time aiming at producing a specific quality. The qualities were described in terms of laver's articulatory model [7]. The experiment is very interesting, because it requires the speakers to systematically explore their whole vocal abilities. Moreover, as the qualities are defined in articulatory terms, one may hope that it becomes possible to correlate specific LTAS changes with specific behaviours, described in a physical or anatomical way. Nolan performed a perceptual analysis of the spectral traces and used several quantitative techniques to characterize each spectral shape by one number (energy ratios, slopes). He nevertheless did not quantify the overall between-spectra differences by means of (dis)similarity measures. It is therefore very difficult to evaluate to what extent the observed changed are important, with respect to the voice quality natural variability.

The aim of our experiment is precisely to combine Nolan's approach, allowing to control the voice variability, with a quantitative technique allowing to measure the resulting spectral variability.

2. EXPERIMENT

2.1. CORPUS

The corpus was a balanced French text, introduced by Harmegnies [8, p.153 : long text A].

One of us (J.E.) produced the text 31 times in succession, managing to give each utterance a specific quality. The qualities he aimed at were the same as in Nolan's experiment, i.e. :

0 neutral
1 raised larynx
2 lowered larynx
3 labial setting with spread lips

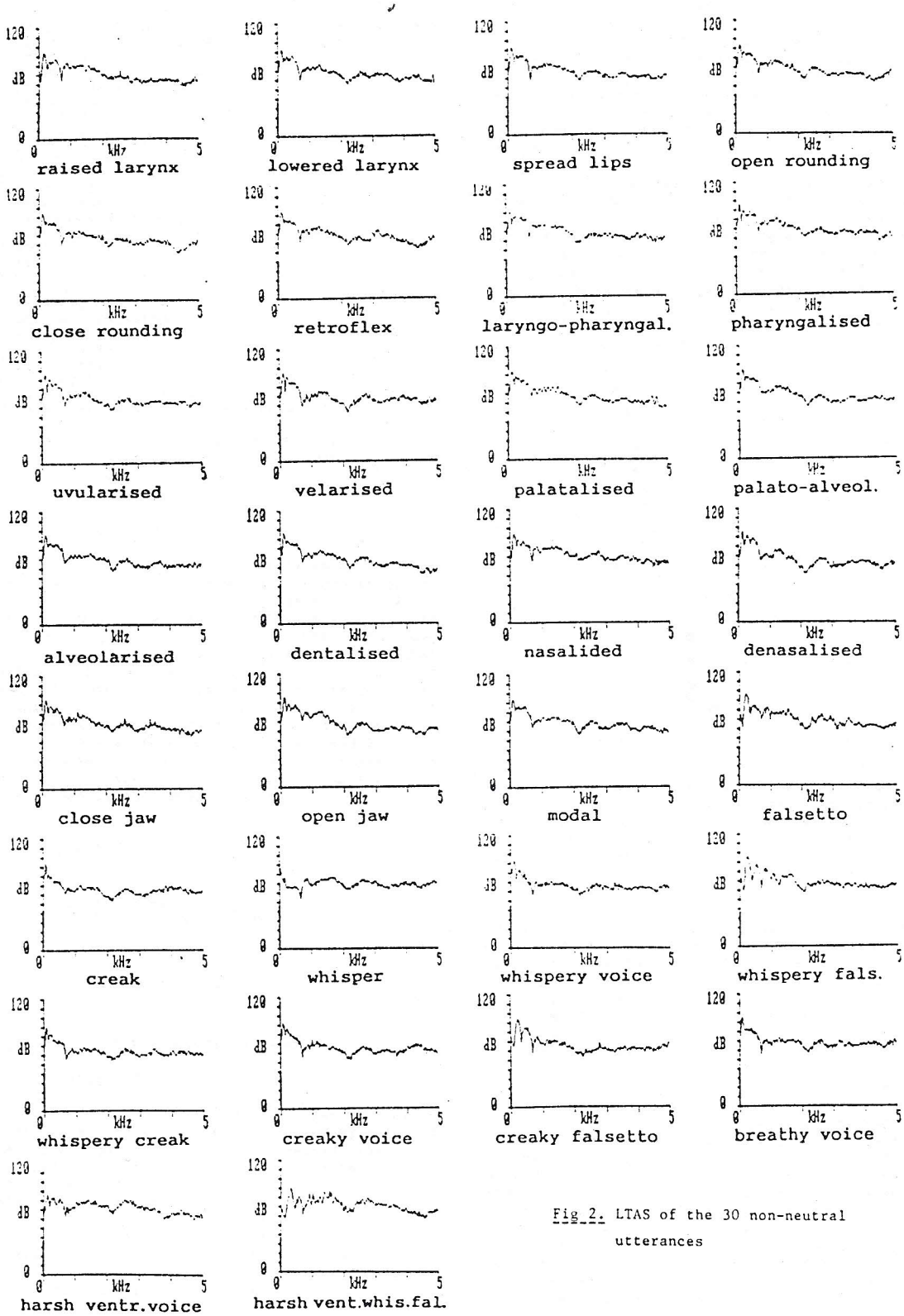

Fig 2. LTAS of the 30 non-neutral utterances

4 labial setting with open rounding and protrusion of the lips
5 labial setting with close rounding and protrusion of the lips
6 retroflex
7 laryngo-pharyngalised
8 pharyngalised
9 uvularised
10 velarised
11 palatalised
12 palato-alveolarised
13 alveolarised
14 dentalised
15 nasalised
16 denasalised
17 close jaw setting
18 open jaw setting
19 modal
20 falsetto
21 creak
22 whisper
23 whispery voice
24 whispery falsetto
25 whispery creak
26 creaky voice
27 creaky falsetto
28 breathy voice
29 harsh ventricular voice
30 harsh ventricular whispery falsetto

The utterances were recorded on a high quality equipment, in a sound-proof room, at the University of Victoria.

2.2. ACOUSTICAL ANALYSIS

The acoustical analysis were later performed at the Département de Phonétique et Psychoacoustique in Mons, by means of a 400 - channel 2033 Bruel Kjaer FFT analyser. Its sampling frequency was set to 12.8 kHz, in order to obtain a DC-5 kHz frequency span. With this setting, the spectra presented a 12.5 Hz resolution over the whole frequency range under investigation. The BK 2033 built-in linear averaging process was used in order to compute 31 LTAS. They are presented in figure 1 and 2. They were afterwards transmitted from the analyser to a 9370 IBM computer via a personnal computer, for storage and further computations.

Fig. 1. LTAS of the neutral utterance.

2.3. COMPARISON TECHNIQUE

Each individual comparison of a given LTAS with another was performed by means of the SDDD index [9], which measures the dissimilarity of the spectra under comparison : high SDDD values indicate considerable dissimilarities between spectra, and conversely.

SDDD is insensitive to variations in the overall levels of the compared spectra, which, therefore, do not need any intensity-normalization prior to the computations.

3. RESULTS

Each LTAS drawn from a non-neutral utterance has been compared with the LTAS computed on the basis of the neutral utterance. The so-obtained values are represented in figure 3. In order to grasp the meaning of these values, it is necessary to compare them with a set of reference values.

Harmegnies [8] has studied the natural residual variability of the LTAS, by measuring how, and to what extent LTAS drawn from successive productions of a single text still vary in the absence of any controllable variations. This experiment involved 1600 LTAS. For comparisons of male voices producing the same text as in the present experiment, he reports SDDD scoring from 1.02 to 4.18. The mean value is 1.88, and the standard deviation .46.

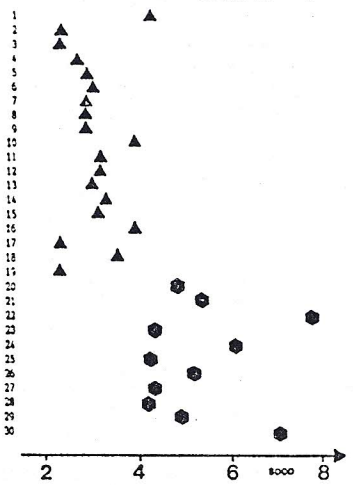

Fig. 3. SDDD values drawn from the comparison of the non laryngeal (triangles) and laryngeal (hexagons) LTAS with the neutral one. The settings are ranked in order of production (see the list in text).

All the values drawn from the present experimentation score higher than the mean of the reference distribution. Thus, as a general rule, it seems that the 30 non-neutral LTAS tend to be more dissimilar from the neutral one than ordinary LTAS are dissimilar one from another. In other words our results seem to confirm that on the whole, the LTAS is sensitive to deliberate changes of voices quality.

It is nevertheless evident that not all the changes exert a major influence on the LTAS. Figure 3 suggests that the computed SDDD values can be divided into two groups. The first one relates to the first 19 settings, and the second, to the remaining 11. Broadly speaking, in the first group, the values are lower than, or approximatevely equall to 4, although in the second, they are higher than 4. It is to be noticed that in the reference distribution [8] the .001 threshold is at 4.08, i.e., 99.9 % of the SDDD values drawn from comparisons of ordinary LTAS, are less than 4.08.

The first group of spectra corresponds to the supralarynge 1 settings. Except for setting 1 (raised larynx), the comparison of these LTAS with the neutral one deliver SDDD values that could belong to the reference intra speaker distribution. Thus, the variations caused by the use of supralaryngeal settings fall within the same range as natural variations.

On the contrary, in the second group, the spectra are computed from utterances using laryngeal settings. The observed SDDD values are incompatible with ordinary intra speaker variations. One may thus conclude that using laryngeal settings induces such major changes in the LTAS shapes that the so obtained spectra are as different from the ordinary ones than spectra computed from different voices are different one from another.

In table 2, the computed SDDD values are presented in order of increasing dissimilarity from the neutral spectrum. The ranking confirms our previous observations, i.e., the last values of the list relate to laryngeal settings. Only one exception occurs : the raised larynx setting, ranked 22. Laver [7] considered this setting supra-laryngeal, because it consists in reducing the length of the supralaryngeal resonators. It is nevertheless obvious that this modification of the voice quality, is due to an action of the larynx, which, in this case, might have influenced the source signal.

Neutral vs lowered larynx	2.28
Neutral vs modal	2.29
Neutral vs spread lips	2.40
Neutral vs close jaw	2.45
Neutral vs open rounding	2.63
Neutral vs pharyngalised	2.66
Neutral vs laryngo-pharyngalised	2.79
Neutral vs close rounding	2.80
Neutral vs uvularised	2.88
Neutral vs alveolarised	2.96
Neutral vs retroflex	2.97
Neutral vs palato-alveolarised	3.16
Neutral vs nasalised	3.18
Neutral vs palatalised	3.20
Neutral vs dentalised	3.38
Neutral vs open jaw	3.64
Neutral vs denasalised	3.96
Neutral vs velarised	4.07
Neutral vs breathy voice	4.30
Neutral vs whispery creak	4.32
Neutral vs whispery voice	4.32
Neutral vs raised larynx	4.35
Neutral vs creaky voice	4.39
Neutral vs falsetto	4.92
Neutral vs harsh ventricular	5.02
Neutral vs creaky falsetto	5.19
Neutral vs creak	5.41
Neutral vs whispery falsetto	6.06
Neutral vs harsh ventr. whisp. fals.	7.24
Neutral vs whisper	7.89

Table 2 : SDDD values drawn from comparisons of the neutral LTAS with the others. The values are ranked in order of increasing dissimilarity from the neutral setting.

The ranking in table 2 moreover suggests that the influence of a given setting on the LTAS depends not only on the organ involved in the modification. The direction of the modification is also important : for instance, the closed and open jaw settings are respectively ranked 4 and 16.

4. CONCLUSIONS

Our results are in good agreement with Nolan's conclusion that the spectrum is more severely perturbed by changes in laryngeal as opposed to supralaryngeal setting. They have moreover led to a ranking of the settings indicating to what extent each of them provokes modifications of the LTAS. A more refined research in this field could lead to the computation of all the inter setting dissimilarities, so that group of settings can be better identified. Finally, It would also be highly desirable to control to what extent subjects effectively perform the articulatory modification they are aiming at (e.g., by means of perceptual examinations and/or of various medical techniques).

5. REFERENCES

[1] BUNGE, E., "Automatic speaker recognition system AUROS for security and forensic voice identification, Proceedings of the International Carnahan Conference on Crime and Countermeasures, 1-7, 1977.

[2] DOHERTY, E.T., "An evaluation of selected acustic parameters for use in speaker identification" , Journal of Phonetics, 4, 321-326, 1976.

[3] FURUI, S., ITAKURA, F., SAITO, S., "Talker verification by long-time average spectrum", Conv. Rec. Acoust. Soc. of Japan, 2-1-2, 1971.

[4] HOLLIEN, H., MAJEWSKI, M., "Speaker identification by long-term spectra under normal and distorted speech conditions", J. Acoust. Soc. Amer., 62, 975-80, 1977.

[5] HOLLIEN, H., MAJEWSKI, W., HOLLIEN, P., "Speaker identification by long-term spectra under normal, stress and disguise conditions", Proceedings of the 8th Int. Congress on Acoustics, 11, 269, 1974.

[6] NOLAN, F., The Phonetic bases of speaker recognition, Cambridge, Cambridge University Press, 1983.

[7] LAVER, J., The phonetic description of voice quality, Cambridge, Cambridge University Press, 1980.

[8] HARMEGNIES, B., Contribution à la caractérisation de la qualité vocale. Analyses plurielles de spectres moyens à long terme de parole, Dissertation doctorale, Université de Mons, Mons, 1988.

[9] HARMEGNIES, B., "SDDD a new dissimilarity index for the comparison of speech spectra", Pattern Recognition Letters, 8, 1988,153-158.

6. ACKNOWLEDGMENTS

The authors wish to thank Mr. Patrick Pairoux for his clever help in processing the data.