ON THE PERCEPTION OF EMOTIONAL CONTENT IN SPEECH

Anne-Maria Laukkanen, Erkki Vilkman**, Paavo Alku*** and Hanna Oksanen*****
** Institute of Speech Communication and Voice Research, University of Tampere,*
***Department of Otolaryngology and Phoniatrics, University of Oulu,*
****Acoustics Laboratory, Helsinki University of Technology,*
***** Department of Health, Faculty of Medicine, University of Tampere,*
Finland.

ABSTRACT

From a nonsense utterance produced expressing neutral state, surprise, sadness, enthusiasm and anger the first 200 ms of the main stress carrying syllable were played at equal loudness to the listeners. Surprise, sadness and anger were especially correctly identified. As the differences in F0 level were artificially eliminated, neutrality, surprise and anger were still identified due to differences in intrasyllabic F0 change and glottal waveform.

INTRODUCTION

The aim of this experiment was to study the perceptual relevance of various speech variables, mainly that of glottal variation in speech, although the elimination of all other variables is in practice very difficult. The speech material was obtained from an earlier study by Laukkanen et al. [1]. There one male and two female subjects produced a nonsense utterance "paappa paappa paappa" simulating five emotional states: neutral, surprise, sadness, enthusiasm and anger. The main stress was given to the underlined syllable. The utterances were 64 - 100 % correctly identified in a listening test. This utterance was chosen since oral pressure during /p/ was used as an estimate of subglottic pressure. Production of emphatic sentence stress simulating various emotional states was regarded as an ideal context to study glottal variation in speech.

The results of the previous study showed that the stressed syllable always had a higher F0 level than the other syllables in the utterance. The emotional states differed from each other in terms of F0 and SP level and change in these parameters as the first and the third (main stress carrying) syllable of the utterance were compared to each other. There was also significant variation in the time based parameters SQ and QOQ [2 and 3, respectively] derived from the acoustically inverse filtered signal. The results from a variance analysis (GLIM) revealed that the glottal variation was more dependent on emotional state than on F0 and SP level alone. Thus the perceptual role of these various parameters needed to be studied further.

MATERIALS AND METHODS

Test 1. The role of syllable length, SPL and intersyllabic F0 and SPL change was eliminated by cutting only the first 200 ms of the main stress carrying (underlined) syllable of the utterance "paappa paappa paappa" to be evaluated. Test 2. The role of F0 level in the samples used in Test 1 was eliminated by artificial pitch modification. In both tests the samples were evaluated by five students of speech science. They answered in a forced choice test, whether the syllables expressed neutrality, surprise, sadness, enthusiasm or anger.

Pitch was artificially modified using a device specially designed for that purpose [4]. The function of the device is based on a digital circuitry for time compression/expansion of the voiced speech segments. The frequency of the voiced segments of a signal is estimated and changed in real time without affecting the signal length. The formant frequencies are changed simultaneously.

Intrasyllabic F0 and A0 (period amplitude) change and the success of pitch modification were studied by calculating F0- and A0 curves from the samples with a microcomputer based signal analysis system ISA (Intelligent Speech Analyser).

Glottal airflow waveform was estimated from the acoustic signal by the IAIF (Iterative Adaptive Inverse Filtering) method [5-6]. Using both synthetic and natural speech the method has been found to yield fairly reliable estimates for voice sources of different fundamental frequencies and phonation types [6].

RESULTS AND DISCUSSION

Figure 1 shows intrasyllabic F0 and A0 changes of the original samples and Figure 2 the results of pitch modification. Figure 3 shows examples of inverse filtered signal waveforms estimated from the part of the syllables where F0 had reached its maximum value.

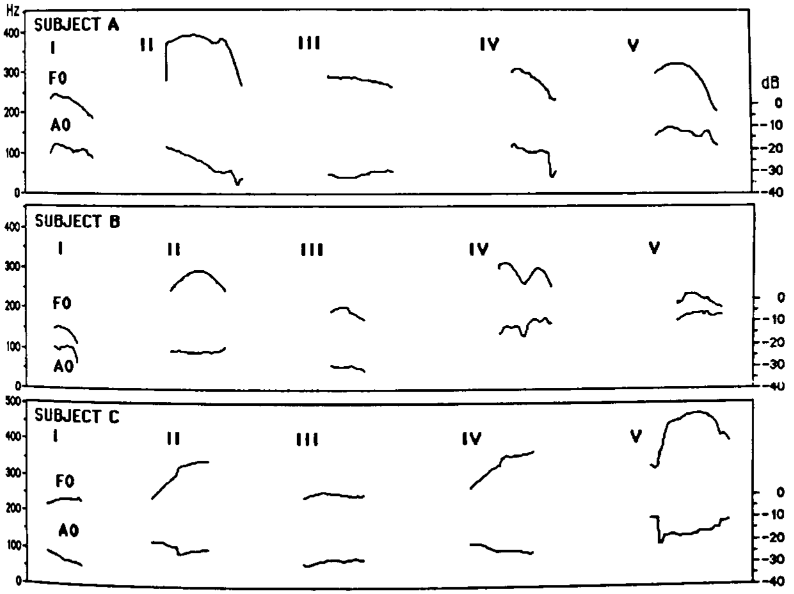

Figure 1. Changes in F0 and A0 during the first 200 ms of the main stress carrying syllable /pa:/. I = neutral state, II = surprise, III = sadness, IV = enthusiasm, V = anger. Subject B is male. Time scale is the same in every sample; differences in the length of the F0/A0-curves is due to the function of the F0 analysis program used.

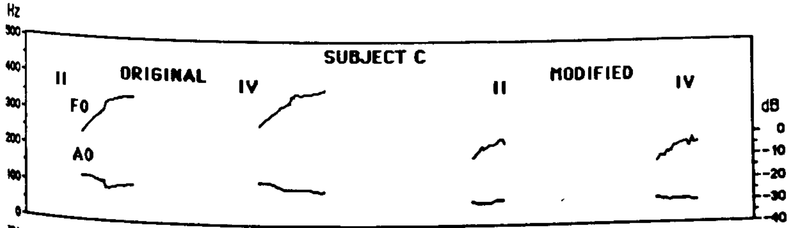

Figure 2. F0 and A0 curves from the original and pitch modified syllables of Subject C (female) expressing (II) surprise and (IV) enthusiasm.

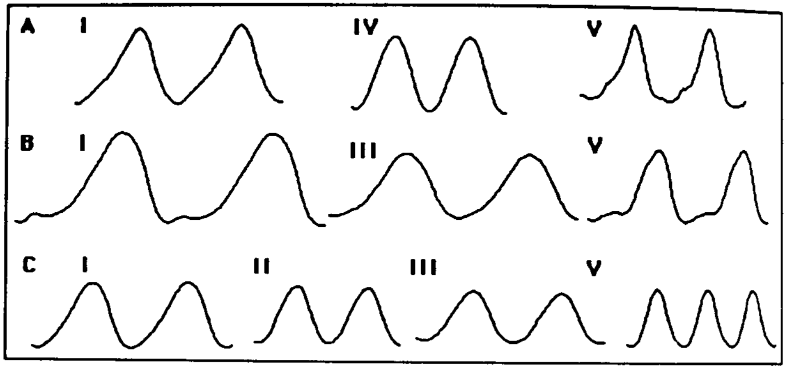

Figure 3. Inverse filtered signal in different samples of Subjects A, B (male) and C. Horizontal axis: time, vertical axis: flow (on an arbitrary scale).

Table 1 shows the results of the listening tests.

Table 1. Number of correct identifications in two listening tests (1. and 2.). Number of listeners = 5. In the first test the first 200 ms of the main stress carrying syllable /paa/ was used. For the second test the same samples were artificially modified in pitch so that all F0 differences between the samples were eliminated. I = neutral state, II = surprise, III = sadness, IV = enthusiasm, V = anger. - = the sample could not be modified.

	Subject		
emotion/test	A	B (male)	C
I 1.	4/5	2/5	3/5
2.	4/5	0	4/5
II 1.	5/5	5/5	4/5
2.	1/5	0	5/5
III 1.	1/5	4/5	4/5
2.	0	3/5	1/5
IV 1.	3/5	0	0
2.	1/5	0	0
V 1.	5/5	5/5	3/5
2.	3/5	5/5	-

In the first listening test the misidentifications were mainly due to the fact that the neutral and sad types were confused, likewise surprise and enthusiasm or enthusiasm and anger. The misidentifications might be based on F0 level, which was lowest in neutral and sad samples and highest in the other samples (Fig. 1). In the second test the number of correct identifications naturally dropped. The neutral samples and those expressing anger were identified best. The misidentifications seemed to emanate from similarities in F0 contour: For Subject A neutral, sad and enthusiastic samples with a falling F0 contour were all identified as neutral. Surprised samples of Subjects A and B as well as the enthusiastic sample of Subject B were misidentified as angry. This may be attributable to the strongly and quickly rising-falling F0 contours in the samples. For Subject B the sad sample was misidentified as surprised, possibly because of the slightly rising contour at the beginning of the sample. This, in turn, may be related to the fact that Subject B seemed to express more compassion than sorrow. For Subject C the enthusiastic sample was misidentified as surprised, most probably because of the similarly rising F0 contours. For Subject C the angry sample could not be pitch modified because of the very high F0 and very pressed voice quality in the original sample.

The subjects differed from each other in their strategies for expression of emotions. For example the fact that in the case of Subjects A and B the number of correct identifications for surprise dropped drastically through pitch modification but in the case of Subject C contrastively the sample was 100 % correctly identified despite the modification suggests that Subjects A and B expressed surprise especially through F0 level but Subject C in contrast through intrasyllabic F0 contour.

A0 changes did not seem to have any independent role in expression, they merely reflected changes in F0.

The most serious disadvantage in pitch modification using this procedure is the fact that in some cases the voice quality becomes unnatural, since the formant frequencies become changed together with F0 and since the digital time manipulation causes some discontinuity of the signal sometimes leading to a slightly clattering sound. However, the results obtained in this experiment were logical suggesting that the quality in the samples was not too much distorted. Furthermore, the pitch modification related formant alteration can be regarded as positive since it eliminates the possible role of formant frequencies in the expression of emotions.

From the bases of this experiment no conclusions can be drawn on the role of the pure glottal airflow waveform in the identification of emotions in speech. The intrasyllabic F0 contour seems to have great perceptual relevance; however, F0 contour naturally also includes dynamic glottal waveform changes. The observations made in this experiment would suggest that voice quality in terms of glottal waveform also had perceptual relevance: In the case of Subject A the enthusiastic sample was by some listeners misidentified as sad, which may be due to the very breathy voice quality in that sample. In the case of Subject C the sad sample was confused almost without exception with the neutral sample, which may be due to the very similar waveforms in both cases (in the period of F0 maximum). In the case of Subject B the F0 contours in the sad and angry samples were fairly similar; however the samples were not perceptually confused, which most likely can be explained by the very different waveforms in them (Fig. 3). However, since the glottal waveform was studied only at F0 maximum, it remains uncertain, whether the possible relevance of voice quality in these samples was related to the glottal waveform in general or to a pitch synchronous change in it. Anyway, the results suggest that also in terms of glottal variation there are individual differences in the strategies for expressing emotions: In these samples Subject C seemed to use less glottal variation and express more through F0 than Subjects A and B (see Figure 3). The role of the glottal waveform in conveying emotions needs to be studied further.

REFERENCES

[1] Laukkanen, A.-M., Vilkman, E., Alku, P., Oksanen, H. (1995). A preliminary study on stress production related physical variations in utterances signalling different emotional states. Submitted for publication.

[2] Timcke, R., von Leden, H. & Moore, P. (1958) Laryngeal vibrations: measurements of the glottic wave. *AMA Arch Otolaryngol.*, 68, 1-19.

[3] Hacki, T. (1989) Klassifizierung von Glottisdysfunktionen mit Hilfe der Elektroglottographie. *Folia phoniatrica* 41, 43-48.

[4] Viitanen J (1982) [Puheäänen korkeuden muuttaminen reaaliajassa]. In Finnish. (Real-time pitch modifying of human voice. Abstract in English). *Folia fennistica & Linguistica*, X Fonetiikan päivät Tampereella 20.-21.3.1981 (Papers from the Xth Meeting of Finnish Phoneticians). Tampereen yliopiston suomen kielen ja yleisen kielitieteen laitoksen julkaisuja 7, Tampere, pp. 323-328.

[5] Alku, P., Vilkman, E. & Laine, U. K. (1991) Analysis of glottal waveform in different phonation types using the new IAIF-method. *Proc. 12th International Congress of Phonetic Sciences, Vol. 4, Aix-en-Provence, 19-24 Aug. 1991*, pp. 362-365.

[6] Alku, P. (1992) Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication* , 11, 109-118.