THE EFFECTS OF EMOTIONS ON VOICE QUALITY

Tom Johnstone and Klaus R. Scherer
University of Geneva

ABSTRACT

Two studies are presented, in which emotional vocal recordings were made using a computer emotion induction task and an imagination technique. Concurrent recordings were made of a variety of physiological parameters, including electroglottograph, respiration, electrocardiogram, and surface electromyogram (muscle tension). Acoustic parameters pertaining to voice quality, including F0 floor, F0 range, jitter and spectral energy distribution were analysed and compared to the physiological parameters. The range of parameters was also statistically compared across different emotions. Although methodological problems still hamper such research, the combined analysis of physiological and acoustic parameters across emotional speaker states promises a clearer interpretation of the effects of emotion on voice quality.

1. INTRODUCTION

In the past two decades, speech scientists have been able to consistently identify important relationships between a number of acoustic speech parameters and speaker attitudes and emotions. On a theoretical level, predictions of emotion-specific acoustic profiles, based on assumptions about physiological components of emotional responses, have met with some empirical success [1,2]. It has been suggested recently that the discrepancies between theoretical predictions and empirical findings may well be due to the almost exclusive use of acted emotional speech in past research, in which actors purposefully adopt highly stereotyped emotion displays that might not reflect the spontaneous, uncontrolled effects of emotion on speech, which were the focus of the predictions made by Scherer [2,3]. A further problem exists in interpreting the data from many studies, due to the fact that discrepancies between predictions and data could have their origins at any of a number of different levels, ranging from the cognitive activity assumed to give rise to emotions, through the physiological responses expected for different emotions, to the way in which vocal characteristics are thought to depend upon speech physiology. Indeed, while much has been said about presumed emotional changes to vocal physiology, empirical research linking physiological measurement with acoustic assessment of emotional speech is almost entirely lacking.

It is increasingly clear that more attention needs to be paid to explicitly linking the cognitive, physiological and acoustic properties of spontaneously elicited emotional speech. Current research by the Geneva Emotion Research Group is pursuing such an approach. A range of methods are being used to elicit naturalistic emotional speech, including computer induction techniques, as well as more classic emotion induction techniques such as those based on imagination. In addition to measurements of the acoustic characteristics of speech, physiological responses are also recorded and analysed.

This paper presents some examples of our recent research. In line with the focus on the uncontrolled, physiologically mediated aspects of emotional speech, acoustic parameters linked mainly to voice quality are examined. These include F0 based parameters as well as measures of spectral energy distribution, as have proved good indicators of expressed emotion in recent research [1,4]. Following on from recent research that has applied inverse filtering of the speech signal to estimate the glottal waveform across different emotions [5,6], we have also started more direct analysis of glottal opening and closing in emotional speech by means of electroglottography. Our research group is also pursuing parallel studies of the physiological responses to emotions that are pertinent to the study of vocal production. The principal aim of such a combined approach is to better understand how the acoustic characteristics of emotional speech are determined by underlying emotional physiological changes.

2. COMPUTER GAME STUDIES OF PHYSIOLOGY AND EMOTIONAL VOICE

In a series of recent and ongoing experiments, we have used a computer game as a means to induce genuine emotional responses in an interactive setting. Selected events in a computer game are systematically manipulated with the intention of eliciting the cognitive appraisals that are postulated by many emotion theorists to underlie emotion.

2.1. Example: the manipulation of goal obstruction and coping potential.

Much evidence in emotion psychology suggests that a person's subjective appraisals of the obstruction caused by an event or stimulus, and their potential to cope with the situation, determine the elicitation of the subsequent emotional response [2,7]. In a recent study, we manipulated the events in a computer space game to provoke appraisals of high or low obstruction and high or low coping potential. Participants were 36 male, native French-speaking students of the University of Geneva, who were competing for three cash prizes (the prizes increase the relevance of the game to the participants, thus rendering the game more emotional). The goal obstruction or conduciveness of game situations was manipulated by the introduction of enemy or friendly alien ships respectively. The coping potential of the player faced with such aliens was manipulated by decreasing or increasing the player's shooting power and spaceship controllability.

During manipulated events, while still playing the game, the participants were asked to provide a report of their emotional state, by pronouncing aloud the phrase "En ce moment, je me sens..." ("At this moment, I feel..."), and then completing the sentence by choosing one or more emotion words from a list of 8 (irritated, disappointed, contented, stressed, surprised, relieved, helpless, and alarmed). These reports were recorded to a Casio DAT recorder using a Sony AD-38 clip-on microphone. Surface electromyographic (EMG) recordings of muscle activity from the forearm extensor of the nonplaying arm, electrocardiogram recordings of heart rate, and respiratory activity measured by means of a thoracic strain gauge were recorded throughout the duration of the experimental session at a sample rate of 800 Hz.

2.2. Analyses

2.2.1. Psychophysiological analyses. Data blocks of length 10 seconds corresponding to each verbal report were extracted for analysis. Each of the EMG signals were rectified and smoothed off-line using a digital Butterworth low-pass filter with a cut-off frequency of 4 Hz., after which the signal was down-sampled to 20 Hz. and averaged over the data block. Mean heart period was calculated by a peak-picking algorithm for the automatic detection of successive R-wave peaks. Respiration rate was calculated by low pass filtering the respiration signal at 4 Hz, and then downsampling the signal to 10 Hz. FFT analysis was then used to approximate the mean respiratory cycle duration

2.2.2. Vocal analyses. Each repetition of "En ce moment, je me sens" was acoustically analyzed using LabSpeech, a suite of semi-automatic LabVIEW speech analysis routines developed in our laboratory[1].

2.3. Results and discussion

Means for the acoustic and physiological parameters for which there were significant main effects of coping or obstruction are shown in Table 1.

Table 1. Estimated marginal means for vocal parameters for which main effects of coping or obstruction were significant.

	coping		obstruction	
	low	high	obstructive	conducive
F0 range (Hz)	38.8	35.4		
Glottal slope			-8.6	-9.2
Heart period (ms)			743	758
Resp. depth	12.0	12.8		
Resp. period (s)	4.0	3.7	4.0	3.8

Acoustic analyses of the vocal recordings revealed effects of player's ability to cope with manipulated game situations on F0 range, with a higher F0 range under low coping conditions than under high coping. The mechanisms behind this difference are unclear. There was a longer respiratory cycle and decreased respiratory depth for low coping situations versus high coping situations, and a similar respiratory cycle effect for obstructive versus conducive events. The players more often reported low coping situations as stressful, which is consistent with the observed shallower respiration. Slower articulation in low coping and obstructive situations due to increased cognitive load might underlie the observed reduced respiratory rate (there was a nonsignificant trend for longer phrase duration in these situations). The increased heart rate (decreased period) for obstructive events supports such an interpretation. In contrast, during high coping and conducive situations, players more often reported being contented or calm, which is consistent with more relaxed speech production and thus normal respiratory depth and rate of articulation. No significant differences between conditions were found for acoustic RMS energy or forearm extensor muscle activity.

It should be noted that all the significant effects reported here were very small. This is probably due in part to the acoustic analyses, which were preliminary and represent coarse-grained averages taken over a whole phrase rather than individual syllables. More detailed analyses are underway. The analysis of respiration was also limited due to a lack of precise synchronisation between the respiratory signal and the acoustic recording.

3. GLOTTAL ANALYSIS OF EMOTIONAL VOICE

In an effort to better understand the link between vocal physiology and the acoustic characteristics of emotional speech, we have recently started to use EGG analysis. The research is motivated largely by the finding that averaged, suprasegmental spectral energy distribution differs significantly across emotions [1]. This effect of emotion on the averaged spectrum might be due in part to differences in the vocal tract, for example caused by changes to precision of articulation and therefore the amplitude and bandwidth of formants. However, in the aforementioned study, emotion effects were found across different spoken sentences, thereby suggesting an articulation-independent cause. Furthermore, in the Banse and Scherer study, as well as more recent unpublished work by our research group, a consistent dependence of the proportion of total energy at relatively low frequencies (i.e. below 1000Hz) on emotion has been observed. Much work on voice source modeling and estimation has pointed to the relationship between the dynamics of glottal flow and the resulting acoustic spectrum, in particular the relative strength of the harmonics [8], suggesting that closer examination of glottal dynamics in emotional speech is merited.

3.1. Example: Glottal analysis of imagined emotions.

This study was primarily carried out as a test of the feasibility and applicability of EGG analysis to the study of emotional voice production. It is intended to apply such analysis to a future study using computer tasks and games to induce emotional speech, as with the previously described study. In this research, eight speakers were asked to imagine themselves in each specific emotional state and then to repeat a number of 5-digit strings and short phrases, as well as the sustained /a/ vowel. Seven non-extreme emotions (tense, neutral, happy, irritated, depressed, bored, anxious) corresponding to those encountered frequently in everyday situations were selected for study. Speech was recorded with a Casio DAT recorder using a Sony AD-38 clip-on microphone to one channel of the DAT. EGG recordings were made with a Portable Laryngograph onto the second DAT channel.

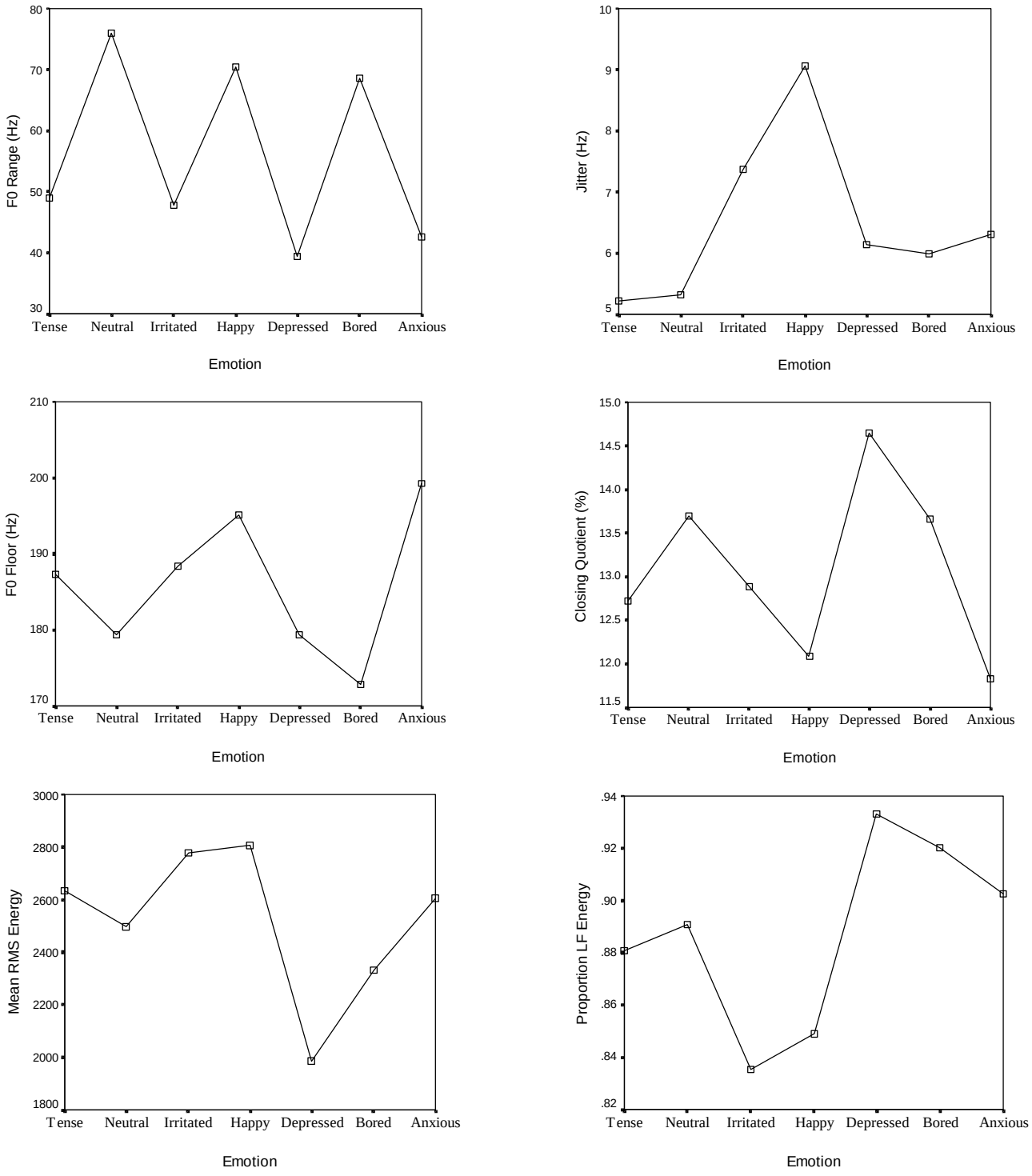

Figure 2. Residual mean values for acoustic and glottal parameters, after the main effects of speaker and utterance have been extracted, shown as a function of emotion (only those parameters for which there was a significant main effect of emotion are shown).

3.2. Analyses.

The samples from each subject were separately acoustically analyzed using LabSpeech. Standard F0 measures were extracted, as well as mean RMS energy of voiced segments and the mean proportion of energy under 1000 Hertz for voiced segments. Opening and closing times, expressed as a percentage of the fundamental period, were calculated from the EGG signal. In addition, a measure of jitter was calculated. For the jitter calculation, a quadratic curve was fitted to a running window of five successive F0 values on the F0 contour and then subtracted from that section of the F0 contour. This served to remove long term, prosodic F0 movements, which would otherwise contaminate jitter measurements. Jitter was then calculated as the mean period to period variation in the residual F0 values.

3.3. Results and discussion.

All acoustic and glottal parameters were tested with

univariate mixed effect ANOVA's, with emotion and phrase as fixed factors and speaker as a random factor. Figure 2 shows the residual mean values for each acoustic and glottal parameter after the main effects of speaker and utterance have been extracted. Only those parameters for which there was a significant main effect of emotion are shown.

As expected, F0 floor was found to be lowest for the emotions bored and depressed, and highest for happy and anxious speech. These results are consistent with predictions of increased muscle tension in high arousal emotions. Consistent with an arousal explanation are the measurements of RMS energy which is low in depressed and bored speech in comparison with the other emotions. The results for F0 range are more difficult to interpret. Depressed speech is seen to have a limited pitch range, which is consistent with prior findings, as is the large range for happy speech. The limited pitch range for tense, irritated and anxious speech might reflect a general tenseness in the laryngeal musculature that limits adjustment of vocal cord tension and larynx position. We have no explanation for the observed high pitch range for bored speech.

The values for jitter are correlated with F0 floor, thus indicating that period to period F0 variation tends to be larger with higher F0. This tendency is absent for anxious and tense speech though, which is in agreement with previous findings of a reduction of jitter for speakers under stress [9].

Analysis of the opening time and closing time of the glottis, expressed as a percentage of T0, showed no significant differences across emotion for opening time, but significant emotion effects on closing time. The variation of the closing quotient across emotions corresponds strongly with the variation of T0 across emotions. Thus for those (high arousal) emotions characterised by high F0 and high RMS energy, the glottis closes faster, as a proportion of the fundamental period. Such rapid closing is a sign of increased vocal effort and/or laryngeal muscular tension as in pressed voice [5,8]. Less damping of harmonics is expected with such phonation, as is indeed indicated by the lower proportion of total energy in low frequency bands for irritated and happy speech. Tense and anxious speech do not show such a reduction in the proportion of low frequency energy however, indicating that the phonation for these emotions is not as easily explained in terms of vocal effort, and that more detailed glottal waveform analysis is warranted.

4. CONCLUSIONS

While there has been an increase in the number of studies concerned with emotional speech, much work is still to be done, in particular with respect to the physiological mechanisms involved in real, rather than acted emotion. Such research often involves practical difficulties with the induction of emotional states in the laboratory, as evident in the first study presented. Although techniques such as the use of computer tasks and games for emotion induction show promise, it is obvious that many difficulties remain. The direct measurement of speech acoustics and speaker physiology, in particular EGG promises more immediate returns. Indeed, in the second study presented here, a relatively simple analysis of glottal opening and closing characteristics proved useful in interpreting the emotion-dependent characteristics of the acoustic waveform. This was despite the fact that analyses were averaged across a number of different utterances, thus showing that such emotion effects are relatively robust to changes in the phonetic context. A more refined glottal analysis of emotional speech based upon the LF-model [8], in combination with speech-synchronised respiratory analysis, is currently being planned.

ACKNOWLEDGMENTS

The authors would like to thank Tanja Banziger and Carien van Reekum for their contributions to speech and physiological recordings and anlysis. This research was funded by the Swiss FNRS grant number 1114-037504.93.

NOTES

1. The routines in LabSpeech are based upon algorithms described in [10]. The software is available in the form of a standalone Windows 95 executable at:
http://www.unige.ch/emotion/members/johnston/software.html

REFERENCES

[1] Banse, R. and Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70(3),* 614-636.

[2] Scherer, K. R. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143-165.

[3] Johnstone, T. & Scherer, K. R. In press. Vocal Communication of Emotion. To appear in: M. Lewis & J. Haviland (Eds.) *Handbook of Emotion* (2nd ed.). New York: Guilford.

[4] Pittam, J., Gallois, C., & Callan, V. 1990. The long-term spectrum and perceived emotion. *Speech Communication, 9(3),* 177-187.

[5] Klasmeyer, G. & Sendlmeier, W. F. 1997. The classification of different phonation types in emotional and neutral speech. *Forensic Linguistics, 4(1),* 104-124.

[6] Laukkanen, A-M., Vilkman, E., Alku, P. & Oksanen, H. 1996. Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics, 24,* 313-335.

[7] Lazarus, R. S. 1991. *Emotion and adaptation*. New York: Oxford University Press.

[8] Fant, G. 1993. Some problems in voice source analysis. *Speech Communication, 13*(1), 7-22.

[9] Smith, G. A. 1977. Voice analysis for the measurement of anxiety. British Journal of *Medical Psychology, 50,* 367-373.

[10] Deller, J. R., Proakis, J. G., & Hansen, J. H. L. 1993. *Discrete-Time Processing of Speech Signals*. New York: Macmillan.