Interactions between Emotional Content of Pictures and Acoustic Features of Speech

R. Ruiz [a] , R. Da Silva Neves [b] , C. Martinot [c], S. Vautier [b]

[a] LA.R.A., [b] C.E.R.P.P., [c] L.D.C.C. : Université de Toulouse-Le Mirail, 5 allées Antonio Machado, 31058 Toulouse cedex 1, France (E-mail: robert.ruiz@univ-tlse2.fr).

Vocal manifestations of picture-induced emotion exist in everyday life. The nonlinguistic information of the speaker's emotional state is investigated here in a laboratory experiment where the correlation between a few acoustic measures on the speech signal and the emotional rating of pictures is studied. The experiment and the first results are presented.

INTRODUCTION

Part of the manifestations of emotion is conveyed by speech. Previous studies have shown that stress and/or emotion modify the speech signal under laboratory and real conditions [1]. Simulation of emotion has been done to study the perception of emotional stress [2, 3]. Various situations have been used to study vocal emotion (psychomotor tasks, cockpit voice recordings, stress tests, etc …). Here, it is induced by pictures. Each one of them is characterized by an "emotional score" composed of an arousal and a pleasure rating between 1 and 9 [4].

A coherent or incoherent text to pronounce is superimposed on the images. And prior to the test subjects have to fill in psychological questionnaires to determine their anxiety state and trait. Therefore multiple correlations are possible not only between acoustical features but also with the emotional scores, the coherence of the sentences and the anxiety.

Time measures performed on the speech signal are studied first. Indeed, analysis of the literature seems to indicate that some of them are the most correlated with various emotionally situations. And unlike the spectral ones they are not phoneme-dependant. They can constitute a reliable indicator basis for the study of emotion detection.

EXPERIMENT AND RESULTS

Pictures belong to the International Affective Picture System (I.A.P.S) database [4]. They are marked by their coordinates in an arousal versus pleasure plane. In four areas of the plane, 24 images are selected (6 by region) : the maximum arousal / maximum pleasure one (A+/P+), the minimum arousal / minimum pleasure one (A-/P-), the A+/P- and the A-/P+ ones. Pictures are proposed in a random order.

Each picture is associated with both a French consistent sentence and an inconsistent one. Each sentence starts by a stop Consonant - Vowel - stop Consonant pseudo-word ($C_1.V.C_2$) corresponding to an element of the picture. For example if there is a river on the screen, the coherent sentence (C.S) is : "tip is a river" (in French) and the incoherent one (I.S) is "tip is a lake" (in French). C.S and I.S are randomly distributed but in equal proportion. C.V.C structures are composed of stop consonants [p], [k], [t]. The choice of stop-consonants is based on the fact that their start is more easily time-detectable than the major part of the other ones because of their impulsive nature. They are associated with vowels [a], [i], [y] in the following combinations : [tip], [pit], [pak], [kap], [tyk], [kyt] . These three vowels are chosen to cover a large frequency domain for future spectral analysis.

A sound signal (like a "bip") is synchronized with image appearance to give the time reference. Indeed, speakers have to pronounce the sentence after they heard the signal. The picture appears during 6 seconds, then the sentence appears during 4 seconds and the "bip" is emitted 6 seconds after the sentence have disappeared from the computer screen.

Before trial pictures, subjects had to utter a list of the six $C_1.V.C_2$ structures used (Phasis 1). In the final phase (Phasis 3) of the experiment, they also have to repeat all the sentences, showed again on the screen, but without the images. These two sets of phonetic material form the pre- and post-"state of rest". During the Phasis 2 the 24 pictures are seen and their associated sentences uttered.

D.A.T recordings are done in a sound studio with an A.K.G cardioid prepolarized condenser microphone (model C420) with a behind-the-neck headband for hands free use. A foam windscreen is used and the microphone is placed near the corner of the mouth to avoid pop noise.

Time measures on the signal are performed by Matlab programs. They are :

1/ the inverse of the mean fundamental period of the low-pass filtered vowel V signal (i.e the mean fundamental frequency) noted $1/T_0$ (in Hz);

2/ the standard deviation of the mean fundamental period computation noted SD_{T0} (in seconds) ;

3/ the jitter of the fundamental period of the low-pass filtered vowel signal (in %).

The first results are reported for only 6 speakers (all students, 3 men and 3 women) and four sets of two pictures both belonging to one of the four groups of the arousal / pleasure plane (Table 1). Table 2 shows the significance of the variations between Phasis 1 (state of rest) and Phasis 2 for all the combinations of the three acoustic features and the four picture groups. For the others comparisons the probabilities of H_0 are of the same magnitude.

Table 1. Results of vocal signal measures are averaged for the 6 speakers studied (standard deviations are into brackets). In the Phasis 2 and 3 results depend on the group the two pictures belong to (A+/P+, A+/P-, A-/P+, A-/P-).

		$1/T_0$ (Hz)	**SD_{T0}** (s)	**Jitter** (%)
Phasis 1		203.16 (65.08)	0.33 (0.24)	4.69 (5.41)
A+/P+	Phasis 2	205.38 (69.15)	0.15 (0.08)	2.74 (1.39)
	Phasis 3	200.52 (67.96)	0.21 (0.12)	4.61 (3.62)
A+/P-	Phasis 2	200.31 (67.93)	0.30 (0.23)	5.20 (3.65)
	Phasis 3	204.92 (68.38)	0.19 (0.12)	3.26 (2.99)
A-/P+	Phasis 2	211.88 (74.09)	0.19 (0.11)	3.16 (2.28)
	Phasis 3	207.81 (71.84)	0.23 (0.18)	4.24 (3.16)
A-/P-	Phasis 2	200.59 (73.61)	0.22 (0.15)	3.47 (2.25)
	Phasis 3	196.59 (72.52)	0.26 (0.21)	4.40 (3.26)

Table 2. Probabilities p of H_0 for the t-test between Phasis 1 and Phasis 2.

	A+/P+	**A+/P-**	**A-/P+**	**A-/P-**
$1/T_0$ (Hz)	0.323	0.322	0.325	0.358
SD_{T0} (s)	0.5	0.492	0.359	0.499
Jitter (%)	0.494	0.453	0.363	0.472

DISCUSSION

Examination of the results leads to the following temporary conclusions.

Numerical variations of the acoustic features mean values between the state of rest (Phasis 1) and the emotional state (Phasis 2) have not of a large extent and are not significant.

Standard deviation of fundamental periods of the segmented vowel and jitter seems to decrease when emotional content is maximum (A+/P+ set of pictures).

Even if the increase of both numbers of speakers and pictures can lead to a better significance of the variations, the characteristics studied will not probably become reliable for the experiment because of the great overlap between the variability inter-speaker and the variations observed. Perhaps the experimental procedure has a masking effect because an important part of the emotion can have disappeared before the "bip" gives the order to speak. A complementary study is needed to ensure that the choice of this reaction time is not too large.

Examination of individual results is more encouraging because largest variations are observed on the speech signal of subjects who seem to less control their emotion before they utter.

CONCLUSION

The experiment needs to be developed with the entire set of images and the twenty five subjects remaining. More investigations about time cues (reaction time, phoneme duration etc …) and spectral cues are also needed with multiple correlation analysis involving the coherence of the sentence and the state of anxiety of the speaker. Results cannot suggest for the moment that vision can acts upon phonation like audition do it. Since the acoustic properties of an utterance are linked to the acoustic properties of a heard sound (Lombard effect for example), they are also probably linked to the visual properties and the emotional content of a viewed picture.

REFERENCES

1. R.Ruiz, E.Absil, B.Harmegnies, C.Legros, D.Poch, "Time- and Spectrum-Related Variabilities in Stressed Speech under Laboratory and Real Conditions," Speech Com., **20** (1-2), 111-129 (1996).
2. I.R.Murray, J.L.Arnott, "Toward the Simulation of Emotion in Synthetic Speech : A Review of the Literature on Human Vocal Emotion, " J. Acoust. Soc. Am., **93**, 1097-1108 (1993).
3. A.Protopapas, P.Lieberman, "Fundamental Frequency of Phonation and Perceived Emotional Stress, " J. Acoust. Soc. Am., **101** (4), 2267-2277 (1997).
4. Center for the Study of Emotion and Attention (CSEA-NIMH) (1999) International Affective Picture System: Digitized Photographs. Gainsville, FL: The Center for Research in Psychophysiology, University of Florida.