CAN WE HEAR SMILE?

Schröder Marc, Aubergé Véronique, Cathiard Marie-Agnès

Institut de la Communication Parlée, ESA CNRS 5009
Université Stendhal/INPG, domaine universitaire
38040 Grenoble Cedex
Tél.: 04 76 82 41 17 - Fax: 04 76 82 43 35
e-mail : auberge@icp.inpg.fr

ABSTRACT

The amusement expression is both visual and audible in speech. After recording comparable spontaneous, acted, mechanical, reiterated and seduction stimuli, five perceptual experiments were held, mainly based on the hypothesis of prosodically controlled effects of amusement on speech. Results show that audio is partially independant from video, which is as performant as audio-video. Spontaneous speech (unvolontary controlled) can be identified in front of acted speech (volontary controlled). Amusement speech can be distinguished from seduction speech.

1. INTRODUCTION

Tartter (11), Tartter & Braun (12) showed that, in American speech, smiling is audible in one-syllable nonsense words. It is a more generally established fact that emotional expressions can be recognized from the auditory signal (1).

The final aim of this work is to synthesize the amusement emotion, both in visual expression – smile – and in prosodic/acoustic features. The existence of different types of smiles were shown by Ekman et al (3), who found a significant difference between an amusement (Duchenne smile), and other smiles without relation to amusement. Ohala (8) distinguishes the expression of an interior emotional state from emotional display having a social signaling function. Provine (9) remarks that most of naturally occurring laughters are not due to amusement, but occur in a positive social context, in many cultures and languages. Damasio (2) describes different brain mechanisms for the involuntary expression of emotion and for the voluntary emotional display.

After a previous work on attitudes (7) our point is mainly to discover if the involuntary and voluntary expression of amusement can be perceived by human in audio-visual productions and modelized for TTS applications. In this first approach of emotions modelling, we retain amusement for many reasons: the concept of amusement seems clear in front of other emotions namings; amusement is expressed both by audio and video media; there are several previous studies on amusement expressions, from smile to laughter (1, 2, 3, 9,11) (this continuum was shown by Fried et al (4): the duration and intensity of expression, from a small smile to a contagious laughter, is related to the degree of stimulation); it appeared a priori possible to build a controlled corpus in order to oppose some chosen features in perceptual experiments and to measure comparable parameters in audio-visual signals.

Thus, we recorded four male speakers in the controlled conditions described hereafter. On these data we held 5 perceptual experiments linked to the five following questions:

- Is amusement audible? Is amusement visible? Experiment (a) (referential for the four following experiments) opposed acoustic amused vs. neutral speech and visual amused vs. neutral speech.

- Are acoustic features of amusement more than the consequences of smile gesture, i.e. a controlled prosody? Experiment (b) opposed acoustic signals produced with spontaneous vs. mechanical smile.

- Is it possible for the speaker to reiterate (in a low-level loop) an "amused" audio-visual signal, as it is the case for acoustic non emotional prosody (?)? Experiment (c) opposed audio-visual spontaneous vs. reiterated speech.

- Is it possible to simulate the expression of the amusement emotion without the causal emotion? Experiment (d) opposed audio-visual spontaneous vs. acted speech.

- Is it possible to distinguish amusement - unvoluntary - smile from seduction - volontary - smile? Experiment (e) opposed audio-visual amused acted vs. seduction acted speech.

2. THE "SMILE" CORPUS

2.1. Spontaneous corpus recording

To generate an unexpected (spontaneous) amusement emotion in speakers (3 professional speakers and a naive one), we made them concentrate on a complex background task which consisted in reading from a screen a "neutral" written utterance, topped by the picture of a face supposed to be a locutor's one, and to reiterate this utterance with a canonical syllable "ma". Speakers were informed to participate to a reiteration audio-visual recording.

After 10 such utterances, we distracted the task, replacing the neutral utterance by an unexpected joke commenting an associated funny picture. Then followed three "normal" utterances, and so on with closer jokes, supposed to be increasingly amusing (on the basis of an unformal request which was supposed to classify the jokes). Lips and eyes were made up in indigo blue, and some marks where painted on the zygomatic major muscles for ulterior analysis. The speakers' face and profile were filmed, and they were recorded in a quiet room.

The speaker reactions to the distracting "amusement" events were quite different depending on the nature of the speaker. The professional tried not to be distracted and to control a "serious" enunciation. Consequently, just a few read sentences were pronounced with an amusement expression, and they produced "amused" speech mainly out of the given corpus utterances, on the contrary the naive speaker produced increasing amused speech during the experiment (and ended in laughter).

2.2. Additional corpus

Each speaker had to choose which part of speech was produced in an "amusement state" (we did not impose smile like an obligatory indice). It was not possible to keep exactly the same sentences for all speakers (however, the 3 sentences of speakers J also belong to the 3 other speakers, the 4th sentence of speaker M belongs to the two others). Then, he pronounced these utterances (a) neutral patterns (without any amusement), (b) with a "mechanical" on neutral patterns (c) as an actor (simulating amusement for the same utterances) (d) first in reiterating the spontaneous amusement utterances after mining of the original stimulus, and second in reiterating these utterances in synchrony with the audio-visual mining of his initial productions (e) with the consign to simulate, a (social) seduction smile, i.e. which could be not be engendered from a amusement state.

For the four speakers, we obtained 184 audio-visual utterances (original and additional simuli).

3. THE PERCEPTUAL EXPERIMENTS

3.1. Experiments organization

20 subjects participated to the five small experiments. These experiments gathered in two kinds of tests: a discrimination task for the three first experiments (judges had to choose the more "amused" stimulus between the two) and identification test for the last two (judges had to choose the nature of a stimulus in a closed choice). For all the tests, subjects could indicate their confidence degree with a binary choice.

3.2. Amused vs. neutral - exp. (a)

Stimuli were presented in three sequential conditions: audio only, video only and audio-video.

In audio only, 84% of pairs were correctly discriminated. As a reference, Tartter (11), in a similar experiment protocol, shown that for pairs composed of mechanical vs. neutral smile, 63% of mechanical smile stimuli were chosen as the most amused. This difference (20%) suggests that the acoustic signal contains some supplementary indices which are not an articulatory consequence of smile gestures and which could be relevant for prosodic control. This hypothesis will be confirmed by experiment (b).

In video only, 95% of pairs were correctly discriminated.

In audio-video, 94% of pairs were correctly discriminated.

These two last high scores validate the corpus which clearly contains amused speech.

A Tukey test showed that the audio-video score is not significantly different from the video score, and the audio score is significantly different from the two others ($p<0.01$). It has be noted an expected speaker effect (see figure 1). For the three conditions (A, V, AV) the global rate for each speaker is 96% for speaker D (the naive one), 89% for speaker J, 97% for speaker P and 72% for speaker M. The identification score for M is inferior to other speakers scores ($p<.01$, Tukey test); J is inferior to P ($p<.05$).

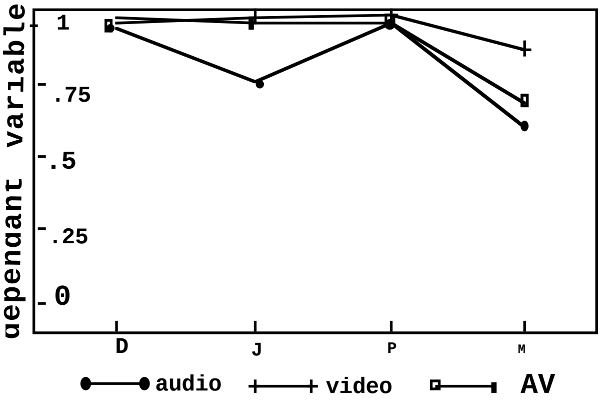

Fig. 1: Identification scores of speakers (D, J, P, M) for amused/neutral pairs globally in the 3 conditions (A, V, AV)

However, we can globally conclude that the visual signal contains enough information to be as performant as the AV signal. Moreover, the acoustic signal seems to have some degrees of freedom in front of video.

3.3. Amused vs. mechanical - exp. (b)

Stimuli were presented in audio only. Global scores confirms the hypothesis of some controlled prosodic events in audio medium: for 69% of pairs, spontaneous stimuli are discriminated as the amused ones. However, figure 2 shows that these results vary according to the speakers (see speaker J for whom scores are haphazard) and to the utterances. It has to be noted that each spontaneous utterance of the corpus contains smiles (with articulatori-acoustic consequences), but often just in a part of the utterance. It was often observed that laugher and amusement smiles generally appear after an utterance. Nevertheless, Provine (9) remarks an hybrid form between speech and laugh, which he says to be consciously controlled. To be more correct, the test could be reiterate with only parts of utterances where spontaneous stimuli contain visual smile.

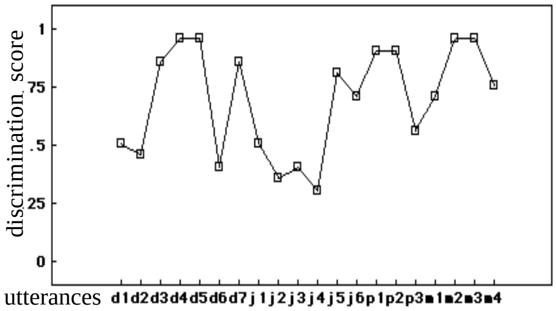

Fig. 2: Scores of discrimination for speakers (D,J, P, M) for each utterance (d1 means speaker D's first sentence ...)

To confirm the part of independance of audio in front of video, the same experiment in video only conditions – not yet held – would be very interesting: results are expected to be the same as for mechanical vs. spontaneous, that means scores are expected to be haphazard.

3.4. Amused vs. reiterated - exp. (c)

Since Liberman et Streeter (6), many works have used or shown the capability to reiterate speech (not only our own speech), in a low-level cognitive loop, with some prosodic characteristics perceived to be similar to the original (see for example Rilliard and Aubergé (10) where a complex reiteration on a canonical syllable is perceived like prosodically identical).

The aim of this test is to observe if such a capability can be performed on emotional speech, in coherence with the hypothesis of a controlled prosody.

To simplify we worked on AV stimuli: the first produced by the speakers after two AV presentations of his own original stimulus, and the second reiterated in synchrony with the AV original stimulus. To be close to preceeding experiments on reiterant speech we should have moreover recorded audio only stimuli after audio miming, to present acoustic pairs to the listeners. At least, we could have performed the test with AV reiterant speech, presented in audio conditions only, to compare the perceptive results to AV conditions. But in this first explanory work we just worked on AV stimuli for the two kinds of reiterations.

In the sequential reiteration task, the original is perceived as the more amused for 56% of pairs, but a z-score approximation shows that the difference when chance is implied is significant. The repartition of scores per sentence is complex. For 1 sentence of D, 2 of P, 1 of Y, 4 of M, original stimuli are identified as more amused with a score of 75%; for 2 sentences of D and 3 of J, on the contrary, the reiterant speech is preferred for 75%; 4 sentences of D, 4 of J, 1 of P, 2 of M, are not significativelly discriminated.

In the synchronous reiteration task, the original is perceived as less amused for 67% of pairs. But for or 1 sentence of D, 4 of J, 2 of M, the choice is significantly at random.

This task seems to be too complex (reiterating in synchronizing both visible gestures and acoustic features) to be performed. But previous anlysis of speech show that timing of acoustic speech and some visible gesture seem well synchronized.

3.5. Spontaneous vs. acted - exp. (d)

Damasio (2) proposes that it is possible to simulate emotion in a loop inside brain , disconnected from the loop inside body which exchanges information with brain during spontaneous emotion. Duchenne de Boulogne in 1862 and then Ekman (3) and Damasio (2) remarked that the expression of a unvolontary amused smile could generate a contraction of the orbicularis oculi (eyes screwing up), which would not be the case of a social (volontary) smile.

Obviously, this experiment strongly depends on the speaker's skill as an actor. But since the spontaneous amusement appeared in very simple and not spontaneous (read) sentences, the differences between speakers, and mainly their absolute actor skill is less important for the task.

In this task, each stimulus (spontaneous and acted sorted randomly) is judged in a closed choice: spontaneous or acted. The correct identification score is 59% (58,3% for spontaneous stimuli and 59,2% for acted stimuli). This score is significantly different from chance (checked by a z-score).

But a variablity between the judges can be observed: 8 subjects (6 females, 2 males) got correct scores between 67% and 90%); 12 subjects got correct scores between 33% and 58%, that means that they answered quite haphazardly.

Moreover the sentences have been globally differentely identified: 2 spontaneous sentences for D, 2 of J, 1 for Y, 3 for M and 3 acted sentences for D, 3 for Y, 1 for M have been well recognized (more than 70%). On the contrary, 2 spontaneous sentences for D, 1 for M and 1 acted sentence for J, 1 for P were bad recognized (more than 70%).

To be interpreted, these results must be detailed and mainly, some visual analysis has to be performed to check for example the Duchenne hypothesis. It would be interesting to identify some specific prosodic cues able to distinguish both volontary controlled vs. unvolontary controlled expression of amusement (i.e. without or with an internal emotional state implied).

3.6. amusement acted/seduction acted/ mechanical smile - exp. (e)

Ohala (8) distinghes two kinds of emotional expressions: one is an unvolontary (but controlled) reaction to the environment, which can or not be decoded by a interlocutor, the other one is volontary produced with the intention of acting on the interlocutor. That is the reason why we selected the seduction expression, which is clearly socially oriented. But since we could not catch some spontaneous seduction smiles from the four speakers and for the same stimuli, we were constrained to use acted stimuli. Consequently, we compare these acted seduction stimuli to acted amused stimuli to neutralize the acting capability effect. Of sure, as we have seen before, we introduced some biais which is the possible Duchenne difference between unvolontray and voluntary smiles.

The identification task consisted in deciding between three possiblities (amusement, seduction or mechanical smile) for each AV stimulus of the randomly sorted stimuli set. Subjects did not know that seduction and amusement were faint.

It has to be noted that the best scores were obtained by the stimuli judged the more complex by the subjects (indicated by a "not sure" choice, and confirmed by the subjects after the experiment).

The mean identification score is 54% (significantly different of 33,3% which is the chance). Table 1 shows the confusion between the different choices. The amusement and mechanical stimuli were recognized the best. The seduction stimuli were half confused with mechanical stimuli, but the mechanical stimuli are not so strongly attracted by the seduction stimuli. This difference is more important related to female vs. male judges, which could perhaps be explained because the speakers are all male (female scores: correct seduction: 42%, confusion with mechanical 42%; correct mechanical 64%, confusion with seduction 18%; male scores: correct seduction: 40%, confusion with mechanical 38%; correct mechanical 54%, confusion with seduction 27%).

	acted stimuli			
scores	amusement	seduction	mechanical	all
amus.	**63,5%**	18,3%	17,9%	34,2%
seduc.	12,7%	**41,2%**	22,4%	25,6%
mecha.	23,8%	40,6%	**59,8%**	40,1%

Table 1. Confusion matrix for the identification test amusement acted vs. seduction acted vs.mechanical smile

However, it cannot be concluded that seduction is difficult to identify. Two classes of stimuli must be separated: clearly bad vs. clearly well recognized. More that half of stimuli was recognized with a score over than 55%, the other half was recognized as haphazard or mechanical. It can be supposed that for these "bad" stimuli, speakers could not express seduction. These results show the necessity to evaluate the quality of the expression. A solution adopted by Banse et Sherer (1) or Leinonen (5) was to preselect the stimuli by experts.

4. CONCLUSION

We can mainly conclude (after other authors), that the amusement expression is greatly audible (84%) compared to not amused stimuli, and we add to Tartter results that the acoustic wave contains some indices specific to the audio medium which could be issued from a prosodic control. The first experiment results were confirmed by the second one: on the base on the acoustic wave, subjects discriminate amused from mechanical stimuli with a 69% score. It would be interesting to complete this experiment in comparing only visually amused vs mechanical stimuli: on this hypothesis, it can be expected both kinds of stimuli cbeing distinguished. On the same hypothesis, with evaluated stimuli, we plan to held a kind of "Mac Gurk" experiment, where seduction, amused and neutral data will be crossed with video and audio to generate some conflictual stimuli.

Since video alone is as performant as audio-video, it will be very interesting to analyze the visual data for an application on visual speech synthesis. At the same time, we will analysis which prosodic indices can be relevant for amusement expression.

Aknowledgment

Many thanks to Christian Abry, who was efficiently generous in giving us many advice and references for our first step in "emotion area".

5. REFERENCES

1. Banse, R., & Scherer, K.R. "Acoustic profiles in vocal emotion expression", *Journal of personality and social psychology*, 170 (3), p. 614-636, 1996.

2. Damasio, A. R. "Descartes' error. emotion, reason, and the human brain." *A Grosset / Putnam books.*, 1994.

3. Ekman P, Davidson, R.J.., & Friesen, W.V. "Duchenne's smile: emotional expression and brain physiology"*J. Pers. Soc. Psych.*, 58, p. 342-353, 1990.

4. Fried, I., Wilson, C. L., MacDonald, K. A., & Behnke, E. J. "Electric current stimulates laughter" *Nature*, 391 (12 february 1998), p. 650, 1998.

5. Leinonen, L., Hiltunen, T., LInnankoski, I., & Laakso, M "Expression of emotional-motivational connotations with a one-word utterance" *Journal of the Acoustic Society of America*, 102 (3), p. 1853-1863, 1997.

6. Liberman M. Y. & Streeter L. A. "Use of nonsense-syllable mimicry in the study of prosodic phenomena" *JASA*, vol 63 (1°, p.231-233, 1978

7. Morlec Y., Bailly G., Aubergé V. "Synthezising attitudes with a global rhythmic and intonation contours" ESCA Workshop on Intonation, 1997.

8. Ohala, J. J. "Ethological theorie and the expression of emotion in the voice" *ICSLP 96*, 1996.

9. Provine, R. R. "Laughter" *American Scientist*, 84 (january-february), p. 38-45, 1996.

10. Rilliard A. and Aubergé, V., The perceptive measure of pure prosody linguistic functions with reiterant sentences, this volume.

11. Tartter, V. C. "Happy talk: Perceptual and acoustic affects of smiling on speech" *Perception & Psychophysics*, 27 (1), p. 24-27, 1980.

12. Tartter, V. C., & Braun, D. "Hearing smiles and frowns in normal and whisper registers" *JASA.*, 96 (4), p. 2101-2107, 1994.