E-WIZ: A TRAPPER PROTOCOL FOR HUNTING THE EXPRESSIVE SPEECH CORPORA IN LAB

Aubergé Véronique, Audibert Nicolas & Rilliard Albert

Institut de la Communication Parlée, Grenoble, France

E-mail: {auberge, audibert, rilliard}@icp.inpg.fr

ABSTRACT

The affects are expressed in different levels of speech: meta-linguistic (expressiveness), linguistic (attitudes), both anchored in the "linguistic time", and para-linguistic (emotions expressions) that is anchored in the timing of the events that cause the emotion. In an experimental approach, the corpora are the basis of analysis. Most of emotional corpus have been produced by acting/elicitating speakers on one hand (with a possible strong control), and on the other hand they have been collected in "real-life". This paper proposes both a Wizard of Oz method for emotional induction and some tools (the E-Wiz platform and a pseudo language-learning software, the Sound Teacher scenario) in order to control the production of authentic data at the paralinguistic level of affects. The proposed protocol has been applied up to now to trap 17 subjects, including actors who reproduced the same utterances with acting methods. The labeling of emotional states was performed by the subjects themselves and is being perceptually validated.

INTRODUCTION

Recent developments of cognitive psychology give to affective processing an increasing role in the cognitive processing. Action tendancy (Frijda, 1987), decision taking (Damasio, 1994): the emotional states variations are proposed as essential evaluation features for the efficiency of cognitive processing. In this frame, the verbal communication needs a coherent affective processing in order to be adapted to the situation. More precisely, it would mean that not to choose to control the affective speech information in synthesis and recognition implies not only naturalness or agreement lacks, but above that perturbs the goals of the interaction themselves. Emotional speech technologies highlight both the questions of psychological emotions representations and expressions modeling, even in the case of stochastic methodologies: as soon as an emotion label is used, some theoretical hypotheses are, even implicitly, implied (e.g. the validity of the "big six" emotions commonly used in speech technologies).

The corpus collection is a key point of the experimental methodologies that are currently used in speech technologies. This paper proposes a way to build authentic corpora for the emotional level of affective speech. After briefly recalling the strengths and weaknesses of *in vivo* vs. *in vitro* methods, E-Wiz, an experimental platform developed at ICP in the frame of the CREST Expressive Speech Project, is presented. A specific E-Wiz application, Sound Teacher, is then more precisely described as well as an example of experiments used to collect very controlled, but however authentic, "pure" direct emotional speech samples. 17 speakers have been recorded in the same induction context, until now for French speakers only, but with the aim to extend the same protocol to English and Japanese.

AFFECTS AND EXPRESSIONS

AFFECTS LEVELS AND EXPRESSIONS FORMS

The affects are expressed in speech through different ways (Aubergé et al., 2003b):

(1) Indirect expression of affects, or expressiveness, is implemented as strategies for the instantiation of linguistic structures. It operates as a meta-control of the linguistic functions of prosody (choice of segmentation size, emphasis, focalization, etc).

(2) Attitudes, i.e. direct expression of the speaker's intentions, voluntarily given by the speaker in addition to the communication purpose and directly encoded as prosodic forms (Aubergé et al., 1997).

(3) The direct expression of the variations of the speaker's emotional states, independently of the communication purpose. Our hypothesis is that this kind of expressions, commonly described as speech expressions, is involuntary controlled by the speaker. The time scale is not anchored in the linguistic events space but in the space of the events that cause the emotions. These are external to the communication context (they can be related by loops links, but anyway considered as external in our view).

The expression stream is generated in parallel to the linguistic and meta-linguistic stream. These two parallel time scales are however integrated in the same speech (prosodic) material. This point is surely decisive in particular to discriminate the communicative vs. para-communicative streams (corresponding for example to the push and pull effects of the Scherer (2001) model).

SPEECH *VS.* FACE EXPRESSIONS

The paradigm of the visible "phonemic" speech (McGurk et al., 1976) shows that a same motor gesture is accessed in both acoustic (sufficient) and visual (partial) modalities. For direct emotion expressions, the face (analog to the speech wave for the phonemic information) is sufficient to give the emotional information. Facial expressions have even been a main basis for emotion theories. The "audible" speech expression is at least a partial consequence of the motor gesture that gives the facial movements. For example, Tarter (1980) has shown that the smile implies some audible deformation of the vocal tract. Many studies have shown that some physiological cues, due to emotion variations, can be heard. But in addition to the bi-modality of a same motor gesture (audible consequences of somatic features), speech also carries some specific information about the emotion variations. Aubergé and Cathiard (2003a) have shown that for the amusement, facially implemented by smile, speech carries some strong information about

emotions that are not only the result of facial smile.

THE SIMULATION PROCESS

Evidence in neurosciences have shown that simulated emotions are produced in the human brain by a simulation loop, i.e. the recollection of the somatic state of a past emotion, whereas authentic emotions result of a cognitive evaluation of stimuli from the whole body (Damasio, 1994). This simulation competence is commonly used in communication (one can simulate to be angry despite he is actually amused), and could be linked to the lying process, processed by specific neural areas. Consequently, the question arises of which expressions are generated in the case of simulation when compared to the expressions of "felt" emotions. Are they fully similar, though voluntary vs. involuntary produced? In particular, what is the time space of such expressions, since there is no event, timed in parallel to the communication stream, that causes the emotion?

Do actors use this simulation competence or not? That is, do they imitate only the output expression (it has been shown for example that a voluntary smile is not controlled in the same neural areas than an amusement smile (Damasio, 1994))? Do they involve the "internal loop", a kind of "emotem", proposed by Damasio (1994)? Or are they able to simulate the whole emotion mechanism in reproducing the complete emotion process including the body?

COLLECTING VOCAL EXPRESSIONS

A SURVEY OF EMOTIONAL CORPORA

A detailed state of the art on emotional samples is given in (Douglas-Cowie et al., 2003). The nature of emotional speech corpus can be classified on three orthogonal dimensions (1) *in vivo vs. in vitro* methods that are laboratory corpora; it implies the recording of speech in good technical conditions (Campbell, 2000); (2) *degree of control* of some of the speech characteristic (interaction situation, linguistic content/generation, phonetic material…); (3) *acted vs. authentic speech* methods.

These dimensions could be wrongly confused because for example authentic data are usually collected *in vivo* outside any control of the observer (that is the one who is collecting). Some authentic data can be collected *in vivo* without the observer's control in very specific contexts (e.g. talk-show corpus (Douglas-Cowie et al., 2000; Chung, 2000)) of banal ecology context (e.g. everyday life corpus for the CREST-ATR team). Very early, some studies have used a control by a partner *in vivo* (e.g. Williams & Stevens (1972) focused on interactions inside a cockpit; Scherer et al. (1984) choose to record social employees in interaction with non professional actors). As for the *in vitro* corpus, recorded in lab, actors have been predominantly used, with less or more sophisticated elicitation method, mainly inspirited from acting methods. The utterances can be linguistically and phonetically pre-defined with or without an emotional content (Amir et al., 1998; Mozziconacci, 1998; Scherer, 2001). Some induction pictures or other kind of stimuli can be presented just before speech production by non-professional speakers (these methods are used in particular to directly observe the somatic changing of emotions when the expressions are not measured). Some non-professional speaker can be asked to read a strongly emotionally loaded talk (Iida, 2002). Data have been collected using actors or non-professional speakers asked to tell a story, develop a theme, describe a past event, which dealt with precise feelings and emotions (Amir et al., 2000; Leinomen et al., 1997). Eventually the stronger control *in vitro* for the collection of authentic data can be obtained by using a partner in precise tasks which precisely constrain the non-emotional speech content and induce some expected emotion variations. Such experiments are much fewer: a computer game situation (Johnstone et al., 1997); an interaction on computer task (Kaiser et al., 1994); a pseudo-phonetic recording (Aubergé et al., 2003a).

AUTHENTIC *VS.* ACTED SPEECH

Acted speech is specific in several points. The first one is that actors produce emotional speech with an artistic goal that can be far from producing speech completely similar to non acted speech (especially in theater methods). The fact that it can be easily identified does not mean that it is identical to non-acted speech. On the contrary, such results can be expected (even acted speech better recognized than non-acted one) for much caricatured, stereotypical acted speech. The second point, already discussed about the simulation process, is that it is impossible to evaluate what the actor imitates. That means that one cannot ensure that acted productions are identical to non-acted ones. In particular, an experiment held by Aubergé and Cathiard (2003a) shows that acted *vs.* non-acted amusement can be discriminated. But more interesting was the fact that some judges' abilities to discriminate were better than others (inter-judge effect), whatever the acting abilities of the speakers could be. Perhaps a very good actor would have avoided such identification and inter-judge effect, but the procedure to evaluate how good an actor is to simulate non-acted speech still has to be developed.

THE WIZARD OF OZ METHOD TO CONTROL DATA

Considering the three bootstrapped levels of affects expressions, it would be particularly interesting to collect the direct emotions expressions in freezing the attitudes and expressiveness variability (ceteris paribus) and to collect the direct attitudinal expressions by freezing the expressiveness. It seems quite impossible to find such representative data in ecological situations. The common way to obtain such a control is to use actors, but, as we noticed before, this it is not an assessable method. Consequently, how to control authentic data production?

Different families of induction scenarios can be proposed, according to types of expressions expected to be collected. In the first case, the subject is convinced to communicate exclusively with a machine, through a very poor and strict word command language, to avoid the use of attitudinal expressiveness, thus restricting the subject's production to direct expressions of emotions. Conversely, in the other kind of scenario, the subject communicates with a human, with common goals to hold a given task on a machine. Again, the command language must not allow any linguistic expressiveness freedom. Finally, the subject uses all linguistic tools for expressiveness in a given task.

Whatever the aim, the data will be authentic but however controlled for their content and recorded in good (*in vitro*)

conditions. It is possible to record different speakers with exactly the same conditions, i.e. one can expect similar reactions. Consequently, it suits well fundamental as well as technological studies.

The 'Wizard of Oz' paradigm, widely used for the evaluation of multimodal interfaces, consists in the imitation by a human partner, called the 'wizard', of the behavior of a complex person-machine interface. The subject believes that he communicates with a computer, whereas the apparent behavior of the application is remote-controlled by the wizard. For the collection of emotional speech corpus, the main interest of that method is to enable the wizard to perturb the application's normal behavior, in order to induce emotional states to subjects. Moreover, it enables to control the phonetic and linguistic contents by the use of a command language that constraints subjects' vocal expression.

The key point to develop such scenarios is to define applications that greatly motivate the subjects: as a matter of fact, their implication is a decisive factor of his reactions to the perturbations, either positive or negative.

E-WIZ: A DEDICATED PLATFORM

In order to set up experiments based on the Wizard of Oz paradigm and aiming at collecting corpus of authentic emotional speech, a dedicated platform, E-Wiz has been developed at ICP. This platform, written in Java language with a client-server architecture (Aubergé et al., 2003b), enables the user to design induction scenarios, without any particular computer-science knowledge. The common frame of such scenarios is to simulate the behavior of a human-machine communication system using voice recognition in order to collect direct emotional expressions in speech. Indeed, the hidden wizard is given the possibility to remote control the application, according to the so-called 'vocal commands' produced by the speaker. The platform is subdivided into three separate applications, including an editor dedicated to the design of scenarios. This editor application aims at generating configuration scripts describing the whole behavior of the client-server applications for a given scenario. Then, a server program running jointly with a client program directly uses those scripts for the actual corpus recording.

Scenarios designed thanks to that software can handle several types of multimedia data, such as texts, images or sounds. Images and texts can be moved by the wizard to produce a kind of slideshow on the client side. In order to facilitate the laying of objects among pages with the editor, particular effort has been made on proposing a user-friendly interface. For instance, editing and word-processing functionalities have been implemented, to enable an intuitive use of the application. Moreover, the task of the wizard may be lightened by making the behavior of some objects automatic. For instance, sounds to be played may be linked with the opening of particular slides, and objects moves may be processed on the client side to seem machine-produced. In addition, automatic countdowns, which behavior when specific values are reached can be predefined, may also be integrated to the slides.

The E-Wiz software is freely available for non-commercial use on request to rilliard@icp.inpg.fr.

THE 'SOUND TEACHER' SCENARIO

E-Wiz scenarios developed for the collection of emotional speech are all based on the same basic principle: subject have to interact with the computer using a command language. The use of a strictly restricted lexicon enables us to collect different emotional expression on the same words, in order to facilitate the acoustic analysis. A first scenario based on logical IQ tests, Top Logic, was first developed but did not motivate the subjects enough.

Sound Teacher is presented as a software enabling the subject to improve his phonetic learning of languages. The subjects are chosen to be strongly motivated by this task. It is supposed to lie on the neuropsychological findings of perception-action theory. It is based on the teaching of 4 vocal tract parameters (opening, front/back, lips rounding, centralization). The subjects are trained to recognize the parameters values when hearing vowels, and to produce them. The scenario is organized in four steps, less to more difficult from the pretext task point of view, and with positive to negative feedback for the Wizard of Oz task. The first step is to check the subject's skills for production and perception of French vowels for French subjects. An artificially positive feedback is given to the subject, quite higher than a supposed averaged score of the others subjects. Then, the subject must learn vowels close to the French vowel system. The feedback is given as higher than the five better performances of preceding subjects. He is informed that his high score enables him to step to a phase of generalization to complex vowels. There, the feedback becomes suddenly negative: the subject is given a score much lower than the average.

He is warned that those results are abnormal, and that his skills for vowels from the French phonological system have to be checked again, since the Sound Teacher software may have perturbed his competences. The last step is thus similar to the first one, but the audio stimuli have been modified to perceptively strongly decrease the vocalic contrasts so that the subject cannot perform the task. He is given scores as the lowest of the preceding subjects. Some commentaries are asked regularly to the subjects, taking as pretext a beta-version of the software.

Each recording session lasted around 50 minutes. For each session, the speech data consist in the command words *'next page'* (in French) repeated 50 to 60 times, and in five monosyllabic words (to avoid timing and long-term prosodic effects) shared in the phonological space ([ʁuʒ], [ʒon], [sabl], [vɛʁ], [bʁik]), repeated 11 to 50 times.

Part of the 17 subjects recorded up to now were professional actors, for which an extra protocol has been used: immediately after having been trapped by Sound Teacher, those subjects were asked to reproduce the expressions of the emotional states they had been encountering during the experiment, using actor's methods. This task was performed both on the utterances used in the spontaneous part and on semantically neutral sentences.

The collected emotions expressed by 17 subjects are close to what was expected: concentration, satisfaction, joy, relief, stress, anger, discouragement, boredom, anguish. It has to be noted that highly coherent groups of reaction appear within subjects, surely linked to their psychological profile. The first emotional labeling is done by the subject himself after the experiment: he is given a VHS video tape, as well

as a pre-filled grid, with the task of describing the different emotional states he has been feeling along the experiment. This labeling is being validated by perceptive tests, as well as the labeling of acted productions.

EXPERIMENTAL MEASUREMENTS

Subjects have been recorded on DAT tape in a soundproof room, with an AKG C1000S microphone, for high quality speech recording. Some references measurements are kept in order to validate the nature, the intensity and the time location of emotional variations expressions:

- visual signal, that is mainly movements of the face and the upper part of the subject's body ;
- bio-physiological signals (heart rate, galvanic skin response, respiration, temperature, electromyography recorded with the Pro-Comp equipment) ;
- the articulatory signals related to voice quality (for now only electroglottographic signal, recorded thanks to the experimental platform EVA2).

These signals can be analysed in parallel to perception measurements. They constitute the main indices of "emotional timing" to determine the instants when the prosodic movements, qualifying the emotion expressions, must be measured.

CONCLUSIONS

This work was motivated by the need to collect some authentic emotional speech corpora, controlled to be (1) representative of each of the considered level of affects expressions for the of analyze emotional expressions independently of attitudinal and expressive variations; (2) similar for each recorded speaker in order to analyze inter-speaker variability; (3) similar for each language to analyze inter-language variability; (4) representative of a large scale of emotions for the parametric characterization and discrimination of expressions.

The first point has to be focused since one of our main hypotheses is that affects are expressed in two parallel flows: expressions anchored in the time domain of emotions, attitudes and expressiveness anchored in the linguistic time domain. Following this hypothesis, to model emotional speech is overall a timing problem that can be solved by analyzing data separately, level by level.

The E-Wiz platform appeared to be an efficient tool to design and implement emotion induction scenarios. The Sound Teacher application enabled us to collect large emotional state variations, from positive to negative values. These data may be used both for the evaluation of voice quality algorithms (from EGG and acoustic signals, (Rossato et al., 2004)), for the study of the acoustical morphology of expressive speech (Aubergé et al., 2004) and to build models of inter-subject, inter-emotion and inter-language variability.

ACKNOWLEDGMENTS

This work is part of the "Expressive Speech Project", held by the CREST/Japan Science and Technology and directed by Nick Campbell. It was done in a close collaboration with Nick's team.

REFERENCES

Amir, N. & Ron, S. (1998). Towards an Automatic Classification of Emotions in Speech. In Proceedings of ICSLP.

Amir, N., Ron, S., & Laor, N. (2000). Analysis of an Emotional Speech Corpus in Hebrew. In Proceedings of ISCA WS on Speech and Emotion, 29-33.

Aubergé, V., Audibert, N. & Rilliard, A. (2003b). Why and how to control emotional speech corpora. In Proceedings of Eurospeech, 185-188.

Aubergé, V., Audibert, N. & Rilliard, A. (2004, to be published). Acoustic Morphology of Expressive Speech: What about Contours? Speech Prosody.

Aubergé, V., Grépillat, T. & Rilliard, A. (1997). Can we perceive attitudes before the end of sentence? In Proceedings of Eurospeech, 2, 871-877.

Aubergé, V. & Cathiard, M. (2003a). Can we ear the prosody of smile? Speech Communication special issue on Speech and Emotion.

Campbell, N. (2000). Databases of Emotional Speech. In Proceedings of ISCA WS on Speech and Emotion, 34-38.

Chung, S (2000). L'expression et la perception de l'emotion dans la parole spontanée. PhD thesis, Université de Paris III.

Damasio, A. R. (1994). Descartes error. Emotion, reason, and the human brain. A Grosset/Putnam Books.

Douglas-Cowie, E., Campbell, N., Cowi, R. & Roach, P. (2003). Emotional speech: towards a new generation of databases. Speech Communication special issue on Speech and Emotion.

Douglas-Cowie, E., Cowie, R. & Schröder, M. (2000). A new emotion database: considerations, sources and scope. In Proceedings of ISCA WS on Speech and Emotion, 39-44.

Frijda, N. H. (1987). Emotions, Cognitive structures and Action tendency. Cognition and Emotion, 1, 115-143.

Iida, A. (2002). A study on corpus-based Speech Synthesis with Emotion. PhD thesis, Keio University.

Johnstone, T. & Scherer, K. R. (1999). The effects of emotions on voice quality. In Proceedings of XIVth ICPhS, 2029-2032.

Kaiser, S. & Wehrle, T. (1994). Emotion research and AI: some theoretical and technical issues. Geneva Studies in Emotion and Communication, 8, 1-16.

Leinomen, L. & Hiltunen, M. L. (1997). Expression of emotional-motivational connotations with one-word utterance. JASA.

McGurk, H. & Mc Donald, J. (1976). Hearing lips and seeing voices. Nature. 264, 746-748.

Mozziconacci, S. (1998). Speech Variability and Emotion : Production and Perception. PhD Thesis, Eindhoven University.

Rossato, S., Audibert, N. & Aubergé, V. (2004, to be published). Emotional voice measurement : a comparison of articulatory-EGG and acoustic-amplitude parameters. Speech Prosody.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking." In K. Scherer, A. Schorr & T. Johnstone (Eds.). Appraisal processes in emotion: Theory, Methods, Research, 92-120: Oxford Uni Press.

Scherer, K. R., Ladd, D. R. & Silverman, K. E. A. (1984). Vocal cues to speaker effect: testing two models. JASA, 76 (5),1346-1356.

Tarter, V. C. (1980). Happy talk: perceptual and acoustic effects of smiling on speech. Perception & Psychophysics, 27 (1), 24-27.

Williams, C. E. & Stevens, K. N. (1972). Emotions and speech: some acoustical correlates", JASA, 52, 4 (2), 1238-1250.