

Why and how to control the authentic emotional speech corpora

Véronique Aubergé, Nicolas Audibert, Albert Rilliard

Institut de la Communication Parlée, UMR CNRS 5009, Grenoble, France

{auberge, audibert, rilliard}@icp.inpg.fr - <http://www.icp.inpg.fr/EMOTION>

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 emotional causes timing. In an experimental approach, the corpus are the base of analysis. Main of emotional corpus have been produced by acting/eliciting speakers on one side (with a possible strong control), and on the other side they have been collected in “real-life”. This paper proposes both to generate a Wizard of Oz method and some tools (E-Wiz and Top Logic, Sound Teacher applications) in order to control the production of authentic data, separately for the three levels of affects.

1. Introduction

The recent developments of the cognitive psychology give to affective processing a more and more central role in the cognitive processing. Action preparing [1], decision taking [2]: the emotional states variations are proposed as essential evaluation features for the efficiency of cognition processing. In such a view, the verbal communication needs a coherent affective processing in order to be adapted to the situation. It would mean in particular that not to choose to control the affective speech information in synthesis and recognition implies not only naturalness or agreement lacks, but above that, it perturbs the goals of the interaction. The interest for affective technologies is increasing [3], as well for the conversational chatterbots (Embodied Conversational Interface Agents of Cassel, Believable Social and Emotional Agents of Bryan Loyall, Affective Agents of Picard...) than for the synthetic speech (see for example proceedings of the ISCA workshop on Speech and Emotion in 2000 and of the 1st JST/CREST International Workshop on Expressive Speech Processing).

Emotional speech technologies highlight both the questions of psychological emotions representations and expressions modeling, even in case of stochastic (not knowledge based) methodologies: as soon as an emotion label is used, some theoretical hypotheses are, even implicitly, implied (see for example the validity/non 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 for speech technologies. This paper proposes a way to build authentic corpus for the different levels of affective speech. After briefly recalling the strengths and weaknesses of *in vivo* vs. *in vitro* methods, is presented E-Wiz, an experimental platform developed at ICP in the frame of the CREST Expressive Speech Project. Some corpora have been collected for attitudes and linguistic speech expressiveness. A specific E-Wiz application, Sound Teacher, is then more precisely presented as an example of experiments in which have been collected very controlled, but however authentic, “pure” direct emotion speech samples. A large number of 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.

2. Affects and expressions

2.1. Different affects levels - different expressions forms

The affects are expressed in speech through different ways:

- (1) The more sophisticated, from a cognitive point of view, is the indirect expression – the expressiveness. It is implemented as the strategies of instantiating the linguistic structures. It means that this way directly concerns the communication purpose. It operates as the meta-control of the linguistic functions of prosody (choice of segmentation size, of the emphasis, focalization etc).
- (2) The direct expression of the speaker intentions, that is his attitudes. This information about the the speaker’s point of view is voluntary given by the speaker as added to the communication purpose. In our model, it is supposed to be directly encoded as prosody forms [4] as direct control on linguistic segments.
- (3) The direct expression of the variations of the speaker’s emotional states during his speech communication context. It does not concern the communication purpose (it can anyway but this is not the expression goal). Our hypothesis is that this kind of expressions, those commonly described as speech expressions, are involuntary controlled by the speaker. The time scale is not anchored in the linguistic events space but in the space of the emotion causes events. 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 distinguish the communicative vs. para-communicative streams (corresponding for example to the push and pull effects of the Scherer [5] model).

2.2. Speech vs. face expressions

The paradigm of the visible “phonemic” speech [6] shows that the 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 for giving the emotion information. Face expressions have been even a main basis of the emotion theories. The “audible” speech expression is at least a partial consequence of the motor gesture that gives the facial movements. For example, Tarter [7] 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 eared. But over the bi-modality of a same motor gesture, the audible consequences of somatic features, the speech carries some specific information about the emotion variations. Aubergé and Cathiard [8] have shown that for the amusement, facially implemented by smile, speech carries some strong information about emotions that are not the result of facial smile.

2.3. The simulation process

Some evidences 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 emotion result of a cognitive evaluation of stimuli from the whole body [2]. This simulation competence is commonly used in communication (we can simulate to be angry where we are amused). It could be supposed to be linked with the lying processing that is treated in specific neural areas. In such a case, which expressions are generated in comparison with the expressions of “felt” emotions? Are they completely similar, since in they 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 expression.

Do the 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 a amusement smile [2])? Do they involve the “internal loop”, a kind of “emotem”, proposed by Damasio [2]? Or are they able to simulate the whole emotion mechanism in reproducing the complete emotion process including the body?

3. Collecting speech expressions

3.1. A survey of emotional corpus

A detailed state of the art on emotional samples is given in [9].

The nature of emotional speech corpus can be classified on two orthogonal dimensions (1) *in vivo* <-> *in vitro* methods that are laboratories corpus; it implies both the recording of speech in good technical conditions (quiet room...) (2) *control* <-> *no control* of some of the speech characteristic (interaction situation, linguistic contain/generation, phonetic material...) (3) *acted* <-> *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 context (e.g. talk-show corpus [10]; [11]) of banal ecology context (e.g. today life corpus for the CREST-ATR team). Very early, some studies have used a control by a partner in vivo in real situations (e.g. Williams & Stevens [12] who focus on interactions inside a cockpit; Scherer et al. [13] chose to record social employees in interaction with non professional actors). For the in vitro – laboratory – corpus, the actors have been predominantly used, with less or more sophisticated elicitation method, mainly inspired from acting methods. The utterances can be linguistically and phonetically pre-defined with or without an emotional contents [14, 5, 15]. Some induction pictures or other kind of stimuli can be presented just before speech producing 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 emotioning talk [16, 17]. Some data have been collected in using actors or non-professional speakers with the task to tell a story/to develop around a theme/ to describe a remember, which have to do with precise feelings and emotions [16, 18, 19]. Finally the stronger control in vitro to collect authentic data can be obtained by using a partner in precise tasks which precisely constrained the non emotional speech contain and induce some expected emotion variations. Such experiments are fewer: a computer game situation [20]; an interaction on computer task [21]; a pseudo-phonetic recording [8].

This work is devoted precisely to collect this kind of data, for the three levels of affects (expressiveness, attitudes and emotions). The aim is to “isolate” as soon as possible the productions of each level in order to analyze them separately.

3.2. Do we need authentic data?

The acted speech is specific in several points. The first one is that actors produce their performances 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 better than non acted) for very caricatured, stereotypical, acted speech. The second point, as already evocated about the simulation process, is that there is no possible evaluation of what the actor imitates. That means that there is no insurance that the acted productions are identical to the non acted productions. In particular, an experiment held by Aubergé and Cathiard [8] shows that acted can be discriminated from non acted amusement. But more interesting was the fact that some of the judges were better to discriminate (inter-judge effect) whatever what the actor abilities of the speakers. Perhaps a very good actor would have avoided such identification and inter-judge effect, but the procedure to evaluate how good is an actor to be similar to non acted speech has to be developed.

3.3. The Wizard of Oz method to control emotional 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 attitudes expressions in freezing the expressiveness. It seems quite impossible to find such representative data in ecological situations. The common way to control data in such a way is to use actors, but it was seen here before, that 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 contain and recorded in very good – in vitro – conditions. It is possible to record different speakers in exactly the same conditions, that is with reactions expected are similar, and to repeat the experiment in different languages. It is thus well adapted for fundamental as well for technological studies.

The “Wizard of Oz” paradigm, widely used for the evaluation of multimodal interfaces, consists in the imitation by a human partner, called “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 the control of 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 greatly motivating the subject: the implication of the subject is a decisive factor of his reactions to the perturbations, either positive or negative.

4. 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 [22, 23], enables the user to graphically design new induction scenarios, without any particular computer-science knowledge. The common frame of such scenarios is to simulate the behavior of 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 (see on the web site).

5. E-Wiz scenarios

5.1. Expressions collecting

The scenarios actually developed in order to collect emotional speech are all based on the same basic principle: subject have to interact with the computer by using a command language. The use of a strictly restricted lexicon is useful to obtain different emotional expression on the same words, in order to facilitate the acoustic analysis.

5.1.1. Top Logic

A first scenario, “Top logic”, has been developed [24]. It is based on IQ logical tests and composed of five sets of ten questions. For each question, the subject have to fill the presented logical series. The answer mode is to choose one object among four by telling its relative position. The selection of the answer is done thanks to the sentence “*the first / second / third / fourth from the left*” (in French). Emotions variations are inducted in the speaker (1) either by manipulating his performances (that could be performed by using easy or very complex logical tests, by shortening the allowed answer time, or by comparing his score to a hypothetical average one, depending on the emotion aimed), or

(2) by tuning the behaviour of the application: simulating a very slow processing of the answers, or introducing a lot of dysfunctions.

5.1.2. Sound Teacher

Sound Teacher is supposed to be a soft to prepare the client to be better in 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 verify 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 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 generalisation to complex sounds. 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 soft. The collected emotions expressed by 10 subjects (many others are under recording) are close to what was expected (surely related to the psychological subject profile): concentration, satisfaction, joy, relief, stress, anger, discouragement, boredom, anguish. The first labelling is done by the subject himself after the experiment and will be verified by perceptive tests.

The speech data consist in the command words “*next page*” (in French), and in five monosyllabic words (to avoid timing and long-term prosodic effects) shared in the phonological space ([*ruʒ*], [*ʒon*], [*sabl*], [*vɛr*], [*brik*]). The comparison between EEG and the Amplitude Quotient algorithm developed by Mokhtari and Campbell [25] for breathy voice detection is an evaluation of this algorithm..

5.2. Attitudes and expressiveness collecting

A first experiment was held on a flight F16 simulator The pilot and the partner had common flying task to perform. The partner disturbed the interaction by positive and negative flying actions. The language being strictly constrained from a linguistic point of view, the subject could express his changing of affects only through his attitudes – that were optimised – and his emotions expressions, but without any expressiveness tools.

Some applications (mainly in language learning) are under development in order to collect ecological and motivated expressiveness samples, in given to the subjects some part of monologs to produce in the interaction with the teacher partner, through a computer learning platform.

6. The experimental measurements

All the scenarios developed with E-Wiz, have been ran in quiet room. Consequently the acoustic recording of speech is high quality. Some references measurements are kept in order to verify the nature, the intensity and the location of emotional variations expressions:

- the visual signal, that mainly the face movements and the top body (since the subject are seat);
- the bio-physiological signals (heart rate, galvanic skin, respiration, temperature, EMG recorded with the Pro-Comp equipment);
- the articulatory signals related to voice quality (for now only EEG).

These signals can be analysed in parallel to further perception measurements. And they are the main indices of "emotional timing" for identifying when to when the prosodic movements, qualifying the emotion expressions, must be measured.

7. Conclusions

Preliminary to this work is the need to collect some authentic emotional speech corpora, controlled to be (1) representative of each of the three levels of affects expressions in order to analyze emotions expression outside the attitudinal and expressive variation, (2) similar for each recorded speaker in order to analyze inter-speaker variability (3) similar for each language in order to analyze inter-language variability (universals as concerns emotions expressions, language specificities as concerns the attitudes and the expressiveness) (4) a large scale of emotions in order to analyze the parametric characterization and discrimination of expressions.

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

The E-Wiz platform (that is free distributed) has been shown to be an efficient tool to build application of emotion induction. The Top Logic and mainly the Sound Teacher applications could obtain from the subject a strong emotional state variation, from positive to negative values.

These data are now used both to verify the validity of voice quality algorithms (on the base on references signals) and to build some model of the inter-subjects, inter-emotions, inter-languages variability.

New applications are under development to increase the domain of the covered emotions and attitudes. Some text vs. speech corpus will be produced to make appear the degree of freedom of prosody for expressiveness.

8. Acknowledgment

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

9. References

- [1] Frijda, N. H. "Emotions, Cognitive structures and Action tendency". In *Cognition and Emotion*, 1, 115-143, 1987.
- [2] Damasio A. R. *Descartes error. Emotion, reason, and the human brain*. A Grosset/Putnam Books, 1994.
- [3] Campbell, N. "Databases of Emotional Speech". In *Proceedings of ISCA 2000*, Northern Ireland, 34-38, 2000.
- [4] Aubergé V., Grépillat T., Rilliard A. "Can we perceive attitudes before the end of sentence?" *Eurospeech 97*, 2, 871-877, 1997
- [5] Scherer, K. R. "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, 2001.
- [6] McGurk H., Mc Donald J. "Hearing lips and seeing voices." *Nature*. 264, 746-748. 1976.
- [7] Tarter, V. C. "Happy talk: perceptual and acoustic effects of smiling on speech", *Perception & Psychophysics*, .27, 1, 24-27, 1980.
- [8] Aubergé, V. & Cathiard M. "Can we ear the prosody of smile?" *Speech Communication - Speech and Emotion*, 2003.
- [9] Douglas-Cowie E., Campbell N., Cowie R. and Roach P. "Emotional speech: towards a new generation of databases". *Speech Communication - Speech and Emotion*, 2003
- [10] Douglas-Cowie E., Cowie R., Schröder M. "A new emotion database: considerations, sources and scope.", *Speech and Emotion ISCA workshop*,.39-44, 2000.
- [11] Chung, S. *L'expression et la Perception de l'Emotion dans la Parole Spontanée*. PhD. thesis, Université de Paris III, 2000.
- [12] Williams, C. E. & Stevens, K. N. "Emotions and speech: some acoustical correlates", *JASA*, 52, 4 (part 2), 1238-1250, 1972.
- [13] Scherer, K. R., Ladd, D. R., Silverman, K. E. A. "Vocal cues to speaker effect: testing two models". *JASA*, 76 (5),1346-1356, 1984.
- [14] Mozziconacci S. *Speech Variability and Emotion : Production and Perception*. PhD Thesis, Eindhoven University, 1998
- [15] Amir, N., & Ron, S. "Towards an Automatic Classification of Emotions in Speech." In *Proceedings ICSLP 1998*,.1998
- [16] Grichkovtsova I. "Acquisition des émotions/attitudes prosodiques du français par des enfants russes ", *Master's degree report*, 2002.
- [17] Iida A. *A study on corpus-based Speech Synthesis with Emotion*. PhD thesis, Keio University, 2002.
- [18] Leinonen, L. & Hiltunen, M. L. "Expression of emotional-motivational connotations with one-word utterance". *JASA*, 1997
- [19] Amir, N., Ron, S., & Laor, N. "Analysis of an Emotional Speech Corpus in Hebrew.", *Speech and Emotion ISCA workshop*,.29-33, 2000.
- [20] Johnstone T. & Scherer K. R.. "The effects of emotions on voice quality". *XIVth ICPhS*, San Fransisco, 2029-2032, 1999.
- [21] Kaiser S. & Wehrle T. "Emotion Research and AI: some Theoretical and Technical Issues." *Geneva Studies in Emotion and Communic.*, 8, 1-16, 1994.
- [22] Audibert N. "Application du paradigme du Magicien d'Oz au recueil de corpus de parole émotionnelle spontanée.", *Master's degree report*, 2002.
- [23] Rebreyend J. "Réalisation d'une application client/serveur basée sur le paradigme du Magicien d'Oz." *Unpublished internal report*, 2002.
- [24] Fouard A. "Etablissement d'un corpus de parole émotionnelle spontanée base sur le paradigme du Magicien d'Oz", *Master's degree report*, 2002
- [25] Mokhtari P.& Campbell N. "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech", *Speech Inf Proc of the IEICE Transactions on Inf and Syst*, 86-. 3, 574-582, 2003.