# EWiz : contrôle d'émotions authentiques

*Nicolas Audibert, Véronique Aubergé & Albert Rilliard*

Institut de la Communication Parlée
Université Stendhal/INPG/CNRS, Grenoble, France
Mél: {audibert, auberge, 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 (emotional expressions) that is anchored in the timing of emotion sources. In an experimental approach, the corpus constitutes the basis of analysis. Most of emotional corpora have been produced by acting/elicitating speakers on one side (with a possible strong control), and on the other side collected in "real-life". This paper proposes both to generate a Wizard of Oz method and some tools (E-Wiz and its scenarios Top Logic and Sound Teacher) in order to control the production of authentic expressions of direct emotions. This method has been applied up to now to 17 subjects, including actors.

## 1. INTRODUCTION

Les développements récents de la psychologie cognitive ont donné au traitement des affects un rôle de plus en plus central au sein des traitements cognitifs. Que ce soit pour la préparation à l'action [7] ou la prise de décision [5], les variations d'état émotionnel sont décrites comme une composante d'évaluation essentielle pour l'efficacité des processus cognitifs. Dans une telle approche, la communication verbale nécessite un traitement affectif cohérent afin de s'adapter à la situation. Cela implique en particulier que le choix de ne pas contrôler les informations vocales d'ordre affectif dans les systèmes de synthèse ou de reconnaissance de la parole se traduit non seulement pas un manque de naturalité, mais surtout par une perturbation du but de l'interaction. L'intérêt pour les technologies affectives s'est accru, autant pour les agents conversationnels que pour la parole synthétique [4], et les technologies de la parole émotionnelle posent à la fois la question de la nature des représentations psychologiques des émotions et celle de la modélisation des expressions, y compris dans le cas de méthodologies stochastiques : dès lors que l'on utilise un étiquetage émotionnel, certaines théories des émotions sont impliquées, explicitement ou non (voir par exemple la question de la validité des six émotions les plus couramment utilisées dans les technologies de la parole, appelées *Big Six*).

La collecte de corpus est un point-clé des méthodologies expérimentales, actuellement utilisées pour les technologies de la parole. Cet article propose une méthode de construction de corpus authentiques applicable aux différents niveaux de parole émotionnelle. L'étude présentée ici est centrée sur le niveau des expressions directes.

## 2. AFFECTS ET EXPRESSIONS

### *2.1. Niveaux d'affects et formes d'expression*

Les affects sont exprimés de diverses façons dans la parole :

(1) La plus sophistiquée d'un point de vue cognitif est l'expression indirecte, ou expressivité, implémentée comme les stratégies d'instanciation des structures linguistiques. Ainsi, l'expressivité est directement liée au but communicatif, et fonctionne comme le méta-contrôle des fonctions linguistiques de la prosodie (choix de la taille des segments, de l'emphase, de la focalisation, etc.).

(2) L'expression directe des intentions du locuteur, c'est-à-dire ses attitudes. Cette information sur le point de vue du locuteur est donnée volontairement, en plus des buts communicatifs. Dans notre modèle, les attitudes sont supposées être directement encodées comme des formes prosodiques, contrôlées par les segments linguistiques.

(3) L'expression directe des variations d'état émotionnel du locuteur lors de la situation de communication, qui ne sont pas nécessairement congruentes au but communicatif. Notre hypothèse est que ce type d'expressions est contrôlé involontairement par le locuteur. Le timing n'est pas ancré dans l'espace linguistique mais dans celui des évènements qui déclenchent les changements d'état émotionnel, externes au contexte de communication.

Le flux des expressions est généré parallèlement aux flux linguistique et méta-linguistique. Ces deux échelles temporelles indépendantes sont toutefois intégrées dans le même matériel prosodique. Cette distinction est certainement fondamentale pour permettre de discriminer les flux communicatif *vs.* para-communicatif (qui correspondent par exemple aux effets *pull* et *push* du modèle de Scherer [12]).

### *2.2. Expressions vocales vs. faciales*

L'effet MacGurk [11] montre qu'on accède à un même geste moteur à la fois dans les modalités acoustique (suffisante) et visuelle (partielle). En ce qui concerne les expressions directes des émotions, le visage (analogue au signal de parole pour les informations phonémiques) est suffisant pour transmettre l'information émotionnelle, à tel point que les expressions faciales ont constitué la base principale des théories des émotions. Les expressions vocales « audibles » sont au moins partiellement une conséquence des gestes moteurs responsables des mouvements faciaux. Ainsi, Tarter [13] a montré que le geste de sourire implique des déformations audibles du conduit vocal. De plus, de nombreuses études ont montré que certains

indices physiologiques, dus aux variations d'état émotionnel, peuvent être entendus. Mais au delà de la bi-modalité d'un même geste moteur, la parole est porteuse d'informations spécifiques sur les variations d'état émotionnel. Aubergé et Cathiard [2] ont montré que pour l'amusement, qui s'exprime au niveau facial par le sourire, la parole est porteuse d'informations fortes sur les émotions qui ne sont pas le résultat du geste de sourire.

### *2.3. Le processus de simulation*

Les neurosciences ont montré que les émotions simulées sont produites dans le cerveau humain par une boucle de simulation, c'est-à-dire le souvenir de l'état somatique associé à une émotion passée, tandis que les émotions authentiques résultent d'une évaluation cognitive de stimuli provenant de l'ensemble du corps [5]. Cette compétence de simulation, qu'on peut supposer liée au processus de mensonge, est couramment utilisée dans les situations de communication (on peut ainsi simuler la colère pour gronder un enfant alors qu'on est en réalité amusé). Comment distinguer dès lors les expressions simulées des expressions d'émotions « ressenties » ? Sont elles totalement similaires, bien que produites volontairement *vs.* involontairement ? En particulier, la question se pose du domaine temporel dans lequel sont ancrées ces expressions. En effet, les émotions simulées ne peuvent être rattachées à unévénement indépendant temporellement du flux communicatif.

## 3. COLLECTER DES EXPRESSIONS VOCALES

### *3.1. Un tour d'horizon des corpus émotionnels*

Un état de l'art détaillé des échantillons de parole émotionnelle collectés jusqu'alors est présenté dans [6]. La nature des corpus de parole émotionnelle peut être classée suivant trois dimensions orthogonales : (1) les méthodes *in vivo*, avec un faible contrôle expérimental, *vs. in vitro*, c'est-à-dire les corpus construits en laboratoire ; (2) le niveau de contrôle de certaines caractéristiques de la parole recueillie (situation d'interaction, contenu linguistique, matériel phonétique, etc.) ; (3) les méthodes actées *vs.* authentiques.

Ces dimensions peuvent être confondues car par exemple les données authentiques sont habituellement collectées *in vivo* sans aucun contrôle de l'observateur (qui est celui qui collecte les échantillons). Ainsi des données authentiques peuvent être recueillies *in vivo* sans le contrôle de l'observateur dans des contextes très spécifiques (comme par exemple des talk-shows) ou au contraire dans une écologie banale (enregistrement de situations de communication de la vie de tous les jours dans l'équipe CREST/ATR). Très tôt des études se sont basées sur l'intervention d'un complice dans une situation réelle (avec par exemple l'étude par Williams et Stevens des interactions à l'intérieur d'un cockpit [14]). Les corpus *in vitro*, réalisés en laboratoire, ont eu recours presque exclusivement à des acteurs, avec des méthodes d'élicitation plus ou moins sophistiquées, inspirées des méthodes d'acteurs. Les énoncés peuvent ainsi être linguistiquement et phonétiquement prédéfinis avec ou sans contenu émotionnel [12]. Des images choisies pour leur potentiel d'induction (ou d'autres types de stimuli) peuvent être présentées juste avant la production de parole par des acteurs non-professionnels. Ces méthodes sont utilisées en particulier pour observer directement les changements somatiques révélateurs des émotions lorsque les expressions ne sont pas mesurées. On peut également demander à des acteurs non-professionnels de lire un texte à fort contenu émotionnel [8], ou à des locuteurs non-professionnels de développer un thème [10] ou de décrire un souvenir [1], en relation avec des émotions ou sentiments précis. Enfin, le contrôle *in vitro* le plus fort peut être obtenu à l'aide de tâches précises permettant de contraindre fortement le contenu vocal non-émotionnel, et avec l'intervention d'un complice pour induire des variations attendues d'état émotionnel. De telles expériences sont en nombre beaucoup plus restreint : une situation de jeu vidéo [8] ; une interaction sur une tâche informatique [9] ; un pseudo-enregistrement de corpus phonétique [2]. Ce travail est précisément dédié au recueil *in vitro* d'expressions émotionnelles directes, afin d'étudier ce niveau séparément.

### *3.2. Pourquoi des données authentiques ?*

La parole actée est spécifique par divers aspects. Le premier est que les acteurs expriment les émotions avec un but artistique qui peut être fort éloigné d'une production authentique. Le fait que ces productions soit aisément identifiables ne signifie pas qu'elles soit identiques à une production non-actée. Au contraire, de tels résultats peuvent être attendus (avec même de meilleurs scores pour les productions actées *vs.* non-actées) pour une parole actée très caricaturale et stéréotypée. Le second aspect, déjà évoqué à propos du processus de simulation, est qu'il est impossible d'évaluer ce que l'acteur imite. Cela signifie qu'on ne peut s'assurer que les productions actées soient identiques aux productions non-actées. En particulier, une expérience menée par Aubergé et Cathiard [2] montre que l'amusement acté peut être discriminé de l'amusement non-acté. Cette étude a de plus montré que certains juges étaient meilleurs que d'autres pour cette discrimination (effet inter-juges) quelque soient les compétences d'acteur des locuteurs. Un très bon acteur aurait peut-être pu éviter une telle discrimination, mais il n'existe pas de procédure permettant d'évaluer la capacité d'un acteur à produire de la parole similaire à la parole émotionnelle non-actée.

### *3.3. Le paradigme du Magicien d'Oz*

Si l'on considère les trois niveaux imbriqués d'expression des affects, il serait particulièrement intéressant de pouvoir collecter les expressions émotionnelles directes en gelant la variabilité due aux attitudes et à l'expressivité (*ceteris paribus*), et de collecter les expressions directes des attitudes en gelant l'expressivité. Il semble presque impossible de collecter de telles données dans des situations écologiques. La manière courante de contrôler ainsi les données est le recours à des acteurs, mais comme nous l'avons vu plus haut ce n'est pas une méthode fiable. Comment alors contrôler la production de données authentiques ?

Le type de scénario d'induction qui peut être proposé pour recueillir des expressions émotionnelles directes est celui dans lequel le sujet est convaincu de communiquer exclusivement avec une machine, à travers un langage de commandes très strict et limité, pour éviter l'usage de l'expressivité attitudinale. Une

telle méthode permet ainsi de recueillir des données authentiques au contenu contrôlé, enregistrées dans des conditions *in vitro*. Il est possible d'enregistrer plusieurs locuteurs dans des conditions identiques, afin que les réactions attendues soient les mêmes, et de répéter l'expérience pour plusieurs langues. Cette méthode est donc bien adaptée à des études à visées fondamentales aussi bien que technologiques.

Le paradigme du Magicien d'Oz consiste en l'imitation par un complice humain, appelé le « magicien », du comportement d'une interface personne/machine complexe. Le sujet croit communiquer avec un ordinateur, alors que le comportement apparent de l'application est contrôlé à distance par le magicien. Pour la collecte de corpus de parole émotionnelle, le principal intérêt de cette méthode est de permettre au magicien de perturber le comportement normal de l'application, afin d'induire certains états émotionnels chez les sujets. De plus, cela permet de contrôler les contenus linguistique et phonétique grâce à l'usage d'un langage de commandes qui contraint les expressions vocales des sujets.

Le point le plus important pour le développement de tels scénarios est de définir des applications permettant une motivation importante du sujet : l'implication du sujet est en effet un facteur décisif pour ses réactions aux perturbations, positives ou négatives.

## 4. E-WIZ : UNE PLATEFORME DEDIEE

Afin de mettre en œuvre des expériences basées sur le paradigme du Magicien d'Oz et destinées à collecter des corpus de parole émotionnelle authentique, une plate-forme spécifique, E-Wiz, a été développée à l'ICP. Cette plate-forme, développée en langage Java avec une architecture client/serveur [3], permet à un utilisateur ne possédant pas de compétences informatiques particulières d'élaborer de nouveaux scénarios d'induction graphiquement. Le schéma commun de ces scénarios est de simuler le comportement d'un système de communication personne/machine basé sur la reconnaissance vocale, afin de recueillir des expressions émotionnelles directes dans la parole. En effet, le magicien a la possibilité de contrôler l'application à distance et à l'insu du sujet, en fonction des pseudo commandes vocales émises par le sujet. E-Wiz se décompose en trois applications distinctes, dont un éditeur dédié à la mise au point des scénarios. L'éditeur permet de générer automatiquement des scripts de configuration décrivant pour un scénario donné le comportement des applications serveur et client, présentées respectivement au magicien et au sujet. Ces applications utilisent directement les scripts générés par l'éditeur pour la phase d'enregistrement effectif des corpus.

Les scénarios élaborés à l'aide de ce logiciel sont constitués d'un ensemble de pages reposant sur divers types de données multimédia, tels que texte, images et sons. Afin de faciliter la disposition des objets avec l'éditeur, ce dernier a été doté d'une interface conviviale. Ainsi, des fonctionnalités d'édition et de traitement de texte ont été implémentées, permettant une utilisation intuitive de l'application. Le magicien a la possibilité d'agir sur ces objets au cours de l'expérience pour modifier le diaporama présenté au sujet. Sa tâche peut être allégée en automatisant le comportement de certains objets. Par exemple, la lecture des sons peut être liée à l'ouverture de certaines pages, et les déplacements des objets peuvent être recalculés côté client afin de paraître provoqués par la machine. Des compteurs automatiques, dont le comportement peut être prédéfini, peuvent en outre également être intégrés aux pages.

La plate-forme E-Wiz est disponible gratuitement pour un usage non-commercial sur demande auprès de rilliard@icp.inpg.fr.

## 5. LES SCENARIOS D'E-WIZ

Les scénarios développés pour la collecte d'expressions émotionnelles directes sont basés sur le principe d'une interaction avec l'ordinateur par le biais d'un langage de commandes. L'usage d'un lexique strictement restreint permet d'obtenir des expressions émotionnelles différentes sur les mêmes énoncés, ce qui facilite l'analyse acoustique.

Un premier scénario, « Top Logic », a été développé. Ce scénario se base sur des tests de logique de type QI, et se compose de 5 séries de 10 questions. Pour chaque question, le sujet doit compléter la suite logique présentée graphiquement. Le mode de réponse est de choisir un objet parmi quatre en le désignant par sa position, avec la commande : *« Le premier/deuxième/troisième/quatrième en partant de la gauche »*. Les variations d'état émotionnel chez le locuteur sont induites soit en manipulant ses performances (en utilisant des questions très simples ou au contraire très complexes, en raccourcissant le temps de réponse accordé, ou encore en comparant le score du sujet à un hypothétique score moyen, en fonction de l'émotion visée), soit en manipulant le comportement de l'application (en simulant un traitement très lent des réponses, voire des bogues).

Le second scénario, Sound Teacher, se présente comme un logiciel permettant au sujet d'améliorer son apprentissage phonétique des langues étrangères. Les sujets sont choisis en raison de leur grande motivation pour cette tâche. La méthode d'apprentissage est supposée reposer sur des découvertes neuropsychologiques relatives à la théorie de la perception/action. Elle se base sur l'apprentissage de 4 paramètres du conduit vocal (ouverture, position avant/arrière, arrondissement des lèvres, centralisation). Les sujets sont entraînés à reconnaître les valeurs des paramètres lors de l'écoute de voyelles, et à les produire. Le scénario est organisé en 4 étapes, du moins difficile au plus difficile du point de vue de la tâche prétexte. La première étape est de vérifier les capacités du sujet pour la production et la perception de voyelles françaises. Un *feedback* artificiellement positif, nettement supérieur au pseudo score moyen, est donné au sujet, qui doit ensuite apprendre des voyelles proches du système vocalique du français. Le score élevé qui lui est attribué (parmi les 5 meilleurs sujets) lui permet alors de passer directement à une phase de généralisation à des sons complexes. Le *feedback* devient alors subitement négatif : le score attribué au sujet est largement inférieur au score moyen des sujets précédents. Le sujet est averti que ces résultats sont anormaux, et que ses compétences pour les voyelles du français doivent être vérifiées à nouveau, car Sound Teacher pourrait les avoir détériorées. La dernière phase est donc similaire à la première, à la différence que les stimuli audio ont été modifiés afin de diminuer fortement le contraste perceptif entre les paires de stimuli présentés, forçant le sujet à répondre au hasard, et qu'un score

très faible lui est alors attribué. Des commentaires sont en outre demandés régulièrement au sujet. Un scénario particulier a été mis au point pour les sujets acteurs, afin de pouvoir comparer des expressions authentiques *vs.* actées : après avoir été piégés par Sound Teacher, ils avaient l'instruction de reproduire à l'aide de méthodes d'acteurs les émotions ressenties au cours de l'expérience.

Le matériel sonore collecté pour constituer le corpus consiste en la commande *« page suivante »*, et en 5 noms de couleur monosyllabiques (pour éviter les effets de *timing* et de prosodie à long terme) répartis dans l'espace phonologique : [ʀuʒ], [ʒon], [sabl], [vɛʀ], [bʀik]. Les émotions recueillies sont proches de celles attendues (avec des classes de réactions suivant le profil psychologique du sujet) : concentration, satisfaction, joie, soulagement, stress, colère, découragement, ennui, angoisse. Un premier étiquetage émotionnel a été réalisé par les sujets eux-mêmes à la suite de l'enregistrement, qui est en cours de validation à l'aide de tests perceptifs.

## 6. MESURES EXPERIMENTALES

Tous les scénarios développés avec E-Wiz ont été mis en œuvre en chambre sourde, afin d'obtenir un enregistrement acoustique de la parole de haute qualité. 17 sujets ont été enregistrés, permettant de recueillir environ 3400 stimuli authentiques. Outre le signal acoustique, certaines mesures de référence sont conservées afin de vérifier la nature, l'intensité et la localisation des variations d'expressions émotionnelles : (1) le signal visuel, relatif aux mouvements de la face et du haut du corps (les sujets sont en effet assis au cours de l'enregistrement), enregistré au moyen d'un caméscope Mini-DV Sony ; (2) les signaux bio-physiologiques (rythme cardiaque, réflexe galvanique, amplitude respiratoire, température, EMG) enregistrés à l'aide de l'équipement Pro-Comp ; (3) les signaux articulatoires relatifs à la qualité de voix (restreints jusqu'alors au signal EGG, mesuré à l'aide de la station d'acquisition EVA2).

Ces signaux, enregistrés sur des canaux séparés, ont été synchronisés *a posteriori* grâce aux « bips » de synchronisation enregistrés à intervalles de temps réguliers au cours de l'expérience, simultanément sur tous les canaux. Ils peuvent être analysés parallèlement aux résultats perceptifs à venir, et constituent les indices principaux du « *timing* émotionnel » pour identifier les instants auxquels les mouvements prosodiques qui qualifient les expressions émotionnelles doivent être mesurés.

## 7. CONCLUSIONS

Ce travail a été motivé par le besoin de recueillir des corpus de parole émotionnelle authentique, contrôlés pour être : (1) représentatifs du niveau d'expression des affects considéré, en éliminant les variations dues aux attitudes et à l'expressivité, (2) similaires pour chaque locuteur afin de pouvoir analyser la variabilité inter-locuteur, (3) similaires pour chaque langue afin d'analyser la variabilité inter-langues, (4) représentatif d'un large panel d'émotions afin d'analyser la morphologie des expressions émotionnelles. Le premier point doit être souligné car l'une de nos principales hypothèses est que les affects sont exprimés suivant deux flux parallèles : les expressions sont ancrées dans le domaine temporel des émotions, tandis que les attitudes et l'expressivité sont ancrées dans le domaine linguistique. Suivant cette hypothèse, la modélisation de la parole émotionnelle est avant tout un problème temporel qui peut être résolu en analysant des données isolées, niveau par niveau.

La plate-forme E-Wiz (disponible gratuitement) s'est révélée être un outil efficace pour construire et mettre en place des scénarios d'induction d'émotions. Les applications de cette plate-forme, Top Logic et Sound Teacher, ont permis d'obtenir chez les sujets de fortes variations d'état émotionnel, avec des valeurs aussi bien positives que négatives.

## 8. REMERCIEMENTS

Ce travail s'inscrit dans le cadre du projet *Expressive Speech Project*, soutenu par le JST/CREST et dirigé par Nick Campbell. Il a été réalisé en étroite collaboration avec l'équipe de Nick.

## BIBLIOGRAPHIE

[1] N. Amir, S. Ron, & N. Laor. Analysis of an Emotional Speech Corpus in Hebrew, *Speech and Emotion ISCA workshop*, 29-33, 2000.

[2] V. Aubergé & M. Cathiard. Can we ear the prosody of smile? *Speech Communication - Speech and Emotion*, 2003.

[3] V. Aubergé, N. Audibert & A. Rilliard. Why and how to control emotional speech corpora. *8th European Conference on Speech Communication and Technology*, 185-188, 2003.

[4] N. Campbell. Databases of Emotional Speech. *Speech and Emotion ISCA workshop*, 34-38, 2000.

[5] A. R. Damasio. *Descartes error. Emotion, reason, and the human brain.* A Grosset/Putnam Books, 1994.

[6] E. Douglas-Cowie, N. Campbell, R. Cowie & P. Roach. Emotional speech: towards a new generation of databases. *Speech Communication - Speech and Emotion*, 2003

[7] N. H. Frijda. Emotions, Cognitive structures and Action tendency. *Cognition and Emotion*, 1, 115-143, 1987.

[8] T. Johnstone & K. R. Scherer. The effects of emotions on voice quality. *XIVth ICPhS*, 2029-2032, 1999.

[9] S. Kaiser & T. Wehrle. Emotion Research and AI: some Theoretical and Technical Issues. *Geneva Studies in Emotion and Communic.*, 8, 1-16, 1994.

[10] L. Leinomen & M. L. Hiltunen. Expression of emotional-motivational connotations with one-word utterance. *JASA*, 1997.

[11] H. McGurk & J. Mc Donald. Hearing lips and seeing voices. *Nature*. 264, 746-748. 1976.

[12] K. R. Scherer. 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.

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

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