La base de données AVLaughterCycle

Jérôme Urbain[1], *Elisabetta Bevacqua*[2], *Thierry Dutoit*[1], *Alexis Moinet*[1], *Radoslaw Niewiadomski*[2], *Catherine Pelachaud*[2], *Benjamin Picart*[1], *Joëlle Tilmanne*[1] *et Johannes Wagner*[3]

[1]TCTS Lab, Faculté Polytechnique, Université de Mons, Boulevard Dolez 31, 7000 Mons, Belgium
[2]CNRS - LTCI UMR 5141, Institut TELECOM - TELECOM ParisTech,
37/39, rue Dareau - 75014 Paris, France
[3]Institut für Informatik, Universität Augsburg, Universitätsstr. 6a, 86159 Augsburg, Germany

ABSTRACT

In this paper, the AVLaughterCycle database is presented. The database consists in the audio and facial motion capture recordings of 24 subjects watching a funny video. It is the first database of laughter combining these 2 modalities. There are around 1000 laughter episodes in the database, covering a large variety of shapes. The database annotation protocol is meant to not only build a large laughter class grouping all the utterances, but distinguish different kinds of laughter. The database is publicly available for research purposes. It can be used for many laughter studies. For example, in the AVLaughterCycle project it was used to endow a virtual agent with the capability of appropriately joining its conversational partner's laughter.
Keywords : laughter, corpus, facial motion capture

1. Introduction

Le rire est un facteur essentiel des relations humaines. Il a la faculté non seulement de transmettre nos émotions mais aussi, par ses vertus communicatives, d'en susciter chez nos interlocuteurs. Différentes disciplines scientifiques (psychologie, médecine, étude du langage, etc.) s'intéressent au rire afin d'identifier ses causes et mécanismes de production, de décrire ses caractéristiques (par exemple, Bachorowski et al. [1]) ou encore de mesurer ses effets. Les progrès en traitement automatique de la parole suscitent un intérêt croissant pour des systèmes capables d'automatiquement détecter le rire ou de le synthétiser de manière naturelle. Si dans un premier temps seul le signal acoustique était étudié (Truong et van Leeuwen [11], Sundaram et Narayanan [10]), Petridis et Pantic [8] y ont récemment ajouté les mouvements faciaux pour détecter le rire. L'émergence des agents virtuels pousse également à combiner les deux modalités (audio et vidéo) afin de doter les avatars de la faculté de rire.

Pour ces travaux de traitement automatique du rire, les données sont cruciales. Mais le rire, comme tous les signaux émotionnels, est difficile à enregistrer car il n'y a aucune garantie qu'un rire acté soit similaire à un rire spontané. Il convient donc, idéalement, d'utiliser des rires naturels, produits par des locuteurs qui ne sont pas conscients qu'ils sont enregistrés. Peu de bases de données ont été enregistrées de cette manière. Celles qui s'en approchent le plus, quoique les sujets savaient qu'ils étaient enregistrés, sont les bases de données ICSI Meeting Corpus (Janin et al. [4]), uniquement audio, obtenue en plaçant des micros lors de réunions techniques, et AMI Meeting Corpus (Carletta [2]), audiovisuelle, consistant en des interview de citoyens racontant des situations émotionnelles. Dans aucun cas le rire n'était l'objectif principal de l'enregistrement des données.

C'est dans ce contexte que s'inscrit cet article. Il décrit une base de données audiovisuelle de rires, appelée AVLaughterCycle, enregistrée dans le cadre du projet eNTERFACE'09 du même nom. Le but du projet était de faire rire un agent virtuel, Greta (Niewiadomski [7]), de manière naturelle en synchronisant les signaux audio et visuels de rires humains. Le signal acoustique était simplement rejoué, alors que les mouvements faciaux ont dû être adaptés à l'animation et la morphologie de Greta. L'application AVLaughterCycle nécessitait une représentation précise des mouvements faciaux lors du rire, données difficiles à obtenir à partir de la base de données AMI. Nous avons donc décidé d'enregistrer une nouvelle base de données, focalisée sur le rire et combinant le signal acoustique à une capture de mouvements faciaux robuste. Les sujets étaient conscients de l'enregistrement : ils étaient laissés seuls dans une pièce éclairée, face à des caméras et avec des marqueurs placés sur le visage. Nous avons utilisé une méthode d'induction pour amener les sujets à l'état souhaité (le rire).

L'article est organisé comme suit : les outils utilisés pour enregistrer les données sont présentés à la Section 2. Le protocole d'enregistrement est exposé à la Section 3. L'annotation est décrite à la Section 4. La Section 5 est dédiée au contenu de la base de données. Enfin, l'application AVLaughterCycle est résumée à la Section 6, avant la conclusion (Section 7).

2. Outils d'enregistrement

La base de données à été enregistrée à l'aide d'une webcam (25 images par seconde, RGB 24 bits, 640x480 pixels) et d'un micro-casque pour l'enregistrement audio (16kHz, PCM 16 bits) et la diffusion de stimulus sonore. De plus, deux outils particuliers ont été utilisés pour construire la base de données : le logiciel "Smart Sensor Integration" (Wagner et al. [13]) et des systèmes de capture de mouvements faciaux. Ces outils sont présentés ci-dessous.

2.1. Smart-Sensor Integration (SSI)

Ce logiciel permet la lecture synchrone de plusieurs signaux d'entrée (dans notre cas, audio et vidéo) et

propose une interface graphique permettant de soumettre les utilisateurs à une série de stimulus audiovisuels (texte, images, vidéos, etc.) sous forme de pages HTML afin de susciter des réactions. Le logiciel permet également l'annotation des données. Les signaux enregistrés peuvent être analysés automatiquement (en temps réel ou non) à l'aide d'algorithmes définis dans une librairie éditable. Via la librairie Torch3D, le logiciel propose également plusieurs types de classificateurs ($HMMs$, $GMMs$, $kNNs$, etc.) pour entraîner des modèles sur les données annotées. Ces modèles peuvent ensuite être utilisés pour classer de nouvelles données, en temps réel ou non.

2.2. Capture des mouvements faciaux

Comme précédemment expliqué, nous souhaitions disposer de mesures précises des mouvements faciaux du rire. Nous nous sommes orientés vers des systèmes de capture de mouvements utilisant des marqueurs placés sur les sujets, plus robustes que les systèmes n'en utilisant pas. Deux logiciels commerciaux ont été successivement utilisés : ZignTrack et OptiTrack.

ZignTrack [6] est un logiciel bon marché nécessitant 22 marqueurs (autocollants ou points de maquillage) et n'utilisant qu'une seule caméra (notre webcam). Les mouvements ne sont donc pas réellement enregistrés en 3D mais en 2D, à partir de laquelle la 3D est regénérée en utilisant un modèle fixe de morphologie faciale. Cela pose des problèmes de reconstruction du visage en 3D lors de rotations de la tête. De plus, le tracking des marqueurs n'est pas assez robuste pour permettre une extraction automatique des mouvements faciaux du rire : de nombreuses corrections manuelles sont nécessaires.

OptiTrack [5] est un logiciel professionnel utilisant 6 caméras infrarouges disposées de manière semi-sphérique. Au moins 23 réflecteurs infrarouges doivent être collés sur le visage des sujets, en plus d'un bandeau muni de 4 réflecteurs placé sur la tête et servant à mesurer les rotations de la tête. Grâce aux 6 caméras, les mouvements sont captés directement en 3D, sans passer par un modèle morphologique. Le tracking réalisé à l'aide de ce système est très robuste.

3. Protocole

Les sujets étaient des volontaires parmi les chercheurs participant au Workshop eNTERFACE'09 à Gênes (Italie). Au total, 24 sujets ont été enregistrés individuellement. Ils étaient originaires de 11 pays : Belgique, Canada, Corée du Sud, Etats-Unis, France, Grèce, Inde, Italie, Kazakhstan, Royaume-Uni et Turquie. 8 participants (3 femmes, 5 hommes) ont été enregistrés avec une capture de mouvements faciaux via le système ZignTrack. Les 16 autres (6 femmes, 10 hommes) ont été enregistrés avec OptiTrack. L'âge moyen des sujets était de 29 ans (écart-type : 7.3 ans).

A cause des capteurs nécessaires à la capture robuste des mouvements faciaux, il nous était impossible d'enregistrer les sujets à leur insu. Pour obtenir des rires spontanés alors que les sujets se savaient analysés, nous avons décidé de leur montrer une vidéo humoristique d'une dizaine de minutes comprenant une série de clips trouvés sur Internet. Les clips se succèdent sans pause : le rire provoqué par un clip peut donc être influencé (écourté ou renforcé) par le clip suivant. Pour éviter des problèmes de sauvegarde, la vidéo a du être séparée en 3 sessions distinctes d'environ 3 minutes lorsque le système OptiTrack était utilisé.

Avant la séance, les capteurs étaient placés sur le visage du sujet. Celui-ci était alors invité à mettre le micro-casque et s'asseoir devant un écran muni de la webcam. Lorsqu'OptiTrack était utilisé, les 6 caméras infrarouges étaient ajoutées autour du sujet. Une page d'instructions lui était alors présentée, avec les consignes suivantes : le sujet devait se détendre et réagir librement à la vidéo, en faisant toutefois attention à garder sa tête vers l'écran et à ne rien placer entre la webcam et son visage tout au long de l'expérience. Une fois les instructions assimilées, le sujet était laissé seul dans la pièce et l'expérience démarrait. A la fin de la vidéo, une page invitait le sujet à produire un rire acté avant de terminer l'enregistrement. Tous les participants ont donné un accord écrit, signé au terme de l'expérience, d'utiliser leurs données à des fins non commerciales. La base de données est disponible à l'adresse *http://tcts.fpms.ac.be/~urbain*. Elle contient les enregistrements audio et les annotations. Les enregistrements vidéo (webcam) et les vidéos de stimulus peuvent être obtenus sur demande. La capture des mouvements faciaux (25FPS avec ZignTrack, 100FPS avec OptiTrack) y sera ajoutée prochainement.

4. Annotation

La base de données est annotée par une personne à l'aide du logiciel SSI. Un protocole d'annotation hiérarchique a été mis au point afin de distinguer 6 classes principales (silence, parole, respiration, applaudissement, rire et "bruit", qui regroupe les sons n'appartenant pas aux 5 autres classes), tout en donnant la possibilité d'ajouter des précisions à l'intérieur d'une de ces classes, en particulier la classe *Rire* qui est le centre d'intérêt de cette base de données. Dans cette catégorie, des précisions peuvent notamment être apportées sur :

- la structure du rire : en spécifiant s'il contient une seule syllabe ou plusieurs et s'il y a plusieurs sections distinctes séparées par des inhalations audibles, appelées "*bouts*".
- le type de son(s) rencontré(s) : voyelle, nasal, chuchotement, grognement, fredonnement, hoquet, etc.

Chaque segment est assigné à une seule classe principale. En cas d'ambiguïté, le segment est annoté "à écarter" afin d'éviter de détériorer les modèles entraînés sur les classes principales (par exemple lorsqu'un téléphone sonne au milieu d'un rire). Il n'y a pas de restriction sur le nombre de précisions apportées : les sous-classes relatives à la structure du rire sont mutuellement exclusives mais peuvent être combinées à toutes les sous-classes de contenu acoustique, car le type de son peut varier au sein d'un épisode de rire.

L'annotation se fait principalement à l'aide du signal audio. Néanmoins, l'enregistrement vidéo est consulté pour affiner les positions des début et fin des segments de rire, ainsi que pour annoter les rires (quasiment) inaudibles. Un rire se termine fréquemment par une

forte inhalation (Chafe [3]), parfois plusieurs secondes après les exhalations principales. Lorsqu'une telle inhalation, manifestement provoquée par le corps du rire qui la précède, est présente, la fin du segment de rire est placée à la suite de cette inhalation.

5. Contenu

La table 1 présente le nombre d'occurrences des classes principales, hormis la classe silence qui est la classe par défaut. La table 2 présente les occurrences des sous-classes de rire. Les rires actés ne sont pas pris en compte. La base de données contient un millier de rires. La plupart des rires est constituée d'un seul "*bout*", contenant lui-même plusieurs syllabes. Les sons du type "voyelle" sont les plus fréquents, mais concernent moins de la moitié des rires. De nombreux rires ont un contenu nasal, assimilable à de la respiration ou à un fredonnement. Les sujets ne disposaient pas d'interlocuteur au cours de l'expérience, ce qui explique la faible fréquence de la classe "Parole" et, en conséquence, la quasi-inexistence de "speech-laughs" (lorsque le sujet parle et rit en même temps, ce qui module le signal de parole (Chafe [3])).

Table 1: Occurrences des classes principales

Classe principale	Occurrences
Rire	1039
Bruit	267
Parole	186
Applaudissement	93
Respiration	41
A écarter	31

Table 2: Occurrences des sous-classes de rire

Catégorie	Sous-classe de rire	Nombre
Structure	Monosyllabique	185
	Un "bout"	697
	Plusieurs "bouts"	157
Acoustique	Voyelle	453
	Nasal	284
	Respiration	245
	Fredonnement	171
	Hoquet	96
	Grognement	18
	Speech-laugh	20
	Silencieux	95

La durée moyenne d'un rire est de 3.5s (écart-type : 5.3s), incluant l'éventuelle inhalation finale. La figure 1 montre un histogramme de la durée des rires et sa fonction de distribution cumulative. Il y apparaît clairement que la plupart des rires sont assez courts (83% des rires durent moins de 5s). Il ne faut néanmoins pas négliger les rires plus longs, qui représentent 51.4% de la durée totale des rires. Le plus long rire dure 82s.

Le nombre et la distribution des rires est extrêmement variable d'un sujet à l'autre : certains sujets rient très peu, d'autres énormément, certains ont tendance à rire brièvement, d'autres produisent de nombreux longs rires, etc. Les causes potentielles de ces réactions variables au stimulus présenté sont nombreuses : différences culturelles, sensibilités diverses à différents

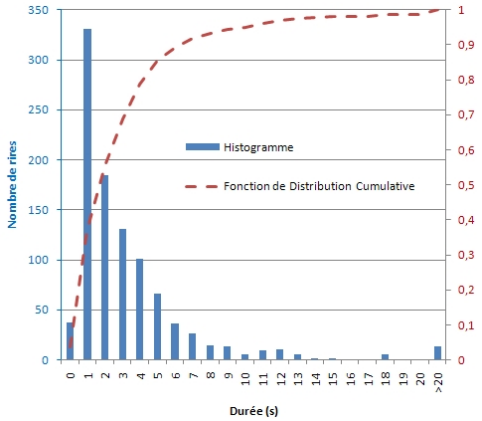

Figure 1: Histogramme et fonction de distribution cumulative de la durée des rires

types d'humour, influence de l'humeur des sujets au moment de l'enregistrement, etc. Des analyses supplémentaires seraient nécessaires pour déterminer les facteurs principaux expliquant les différents comportements constatés.

6. Application AVLaughterCycle

Le but du projet AVLaughterCycle était de construire un système capable d'enregistrer le rire de l'utilisateur et d'y répondre par un rire adéquat. C'est l'agent virtuel Greta qui joue le rire sélectionné. L'architecture de l'application est illustrée à la Figure 2.

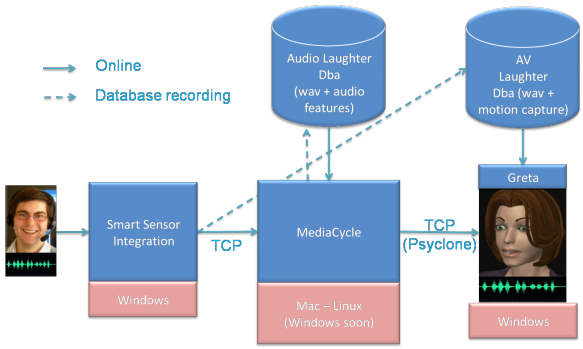

Figure 2: Architecture de l'application AVLaughterCycle

L'utilisateur rit dans le microphone. Le signal audio est analysé en temps réel par le logiciel *SSI* qui en extrait des caractéristiques spectrales pour chaque trame de 340 échantillons, avec un décalage de 85 échantillons entre 2 trames successives. Sur base du rapport signal à bruit, *SSI* segmente le signal audio. Dans le système actuel, il n'y a pas de réelle détection du rire mais une détection d'activité vocale. L'hypothèse est faite que le signal envoyé par l'utilisateur est un rire. Une fois un rire segmenté, *SSI* envoie les moyennes et écart-types de ses caractéristiques spectrales (MFCCs, Spectral Flatness, Loudness, etc.) au deuxième module, appelé MediaCycle.

MediaCycle (Siebert et al. [9]) est un logiciel permettant d'organiser une base de données multimédia et d'y naviguer de manière efficace. La base de données est organisée en fonction des similarités entre les objets. Les similarités sont estimées par la distance Euclidienne entre les vecteurs de caractéristiques des ob-

jets. Pour le projet AVLaughterCycle, seules les caractéristiques spectrales du signal audio ont été utilisées. Pour chaque rire de la base de données, MediaCycle gardait en mémoire les moyennes et écart-types des caractéristiques spectrales, normalisées. Lorsque *SSI* envoie un vecteur de caractéristiques à MediaCycle, celui-ci le compare avec les rires de la base de données. Le rire dont le vecteur de caractéristiques est le plus proche de celui du rire d'entrée est sélectionné. Ce système peut servir à naviguer dans la base de données en l'interrogeant par du rire.

Une fois le rire à jouer sélectionné, sa référence est envoyée à Greta afin qu'elle le joue instantanément : le son n'est pas modifié mais l'animation faciale de Greta est pilotée par les mouvements faciaux du rire, adaptés au visage de Greta (Urbain et al. [12]).

Le projet AVLaughterCycle a débouché sur la réalisation de cette chaîne complète de traitement en quasi temps réel : *SSI* analyse le signal audio en continu et, dès qu'un segment (supposé être du rire) est détecté, le rire le plus similaire est transmis à Greta qui le joue (en plus de son comportement "normal", ses mouvements des bras, etc., qui sont réalisés indépendamment du rire). Le système est opérationnel et les tests qualitatifs qui ont été effectués étaient prometteurs même s'ils ont mis en évidence la nécessité d'introduire des caractéristiques décrivant la structure et le rythme du rire. De plus amples informations sur le projet AVLaughterCycle peuvent être trouvées sur *http ://www.numediart.org/projects/07-4-avlaughtercycle/*. Les résultats de MediaCycle sont actuellement soumises à des évaluations objective (déterminer si MediaCycle est capable de grouper des rires du même type ou du même locuteur) et subjective (mesurer si MediaCycle est en accord avec perceptions de similarité humaines).

7. Conclusion

La base de données AVLaughterCycle, première base de données de rires comprenant à la fois le signal audio et un tracking précis des expressions faciales, a été présentée. Cette base de données est centrée sur le rire et est annotée en fonction, pour donner des indications sur la structure et le contenu de chaque rire. La base de données est disponible gratuitement pour des utilisations non commerciales. Elle contient environ un millier de rires, de types et longueurs variables. Ses applications potentielles couvrent les domaines suivants : l'analyse, la reconnaissance, la modélisation et la synthèse du signal audio du rire ou des mouvements faciaux ; l'étude simultanée de ces deux modalités et leur synchronisation, etc.

Remerciements

Le projet AVLaughterCycle a été partiellement financé par le projet Européen CALLAS (IP6, contrat n° 034800) et par le Ministère de la Région Wallonne en Belgique, via le Programme de Recherche Numédiart (contrat n° 716631). Joëlle Tilmanne dispose d'une bourse doctorale octroyée par le Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.) en Belgique.

Références

[1] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110 :1581–1597, 2007.

[2] J. Carletta. Unleashing the killer corpus : experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2) :181–190, 2007.

[3] W. Chafe. *The Importance of not being earnest. The feeling behind laughter and humor.*, volume 3 of *Consciousness & Emotion Book Series*. John Benjamins Publishing Company, Amsterdam, The Nederlands, 2007.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong-Kong, April 2003.

[5] Natural Point, Inc. Optitrack. http ://www.naturalpoint.com/optitrack/.

[6] Zign Creations. Zign track. http ://www.zigncreations.com/zigntrack.html.

[7] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud. Greta : an interactive expressive ECA system. In C. Sierra, C. Castelfranchi, K. S. Decker, and J. S. Sichman, editors, *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*, pages 1399–1400. IFAAMAS, 2009.

[8] S. Petridis and M. Pantic. Is this joke really funny ? judging the mirth by audiovisual laughter analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1444–1447, New York, USA, June 2009.

[9] X. Siebert, S. Dupont, P. Fortemps, and D. Tardieu. MediaCycle : Browsing and performing with sound and image libraries. In T. Dutoit and B. Macq, editors, *QPSR of the numediart research program*, volume 2, pages 19–22. numediart, 2009.

[10] S. Sundaram and S. Narayanan. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America*, 121(1) :527–535, January 2007.

[11] K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49 :144–158, 2007.

[12] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. AVLaughterCycle : An audiovisual laughing machine. In T. Dutoit and B. Macq, editors, *QPSR of the numediart research program*, volume 2, pages 97–104, Sept. 2009.

[13] J. Wagner, E. André, and F. Jung. Smart sensor integration : A framework for multimodal emotion recognition in real-time. In *Affective Computing and Intelligent Interaction*, 2009.