Workshop Program

1 Preface

Keynotes

3 You had to be there to know why it was funny
Anton Nijholt

5 LoL, Laughing with Greta
Catherine Pelachaud

7 Nonverbal vocalisations are basic human expressions of emotion
Disa Sauter

9 Research on laughter in vocal communication and its progress in the last decade
Jürgen Trouvain

Oral Session

11 Audio-visual laughter synthesis system
Hüseyin Çakmak, Kevin El Haddad, and Thierry Dutoit

15 Towards a speech synthesis system with controllable amusement levels
Kevin El Haddad, Hüseyin Çakmak, Stéphane Dupont, and Thierry Dutoit

19 Using MAGE for real time speech-laugh synthesis
Kevin El Haddad, Alexis Moinet, Hüseyin Çakmak, Stéphane Dupont, and Thierry Dutoit

23 Exploring the body and head kinematics of laughter filled pauses and breaths
Spyros Kousidis, Julian Hough, and David Schlangen

27 The relationship between laughter intensity and perceived humour
Gary McKeown and Will Curran

31 When do children smile? Predicting genuine smiles and non-genuine smiles among winning and losing children
Phoebe Mui, Martijn Goudbeek, and Marc Swerts

35 Laughter detection for on-line human-robot interaction
Marie Tahon and Laurence Devillers

39 Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations
Leimin Tian, Catherine Lai, and Johanna D. Moore

Poster Session

43 With or without words: a pragmatic study of the innovative quotative expressions in English
Sandy Ciroux

45 Interpretation of “embarrassment” laughter in narratives by people with aphasia and non-language-impaired speakers
Mariya V. Khudyakova and Mira B. Bergelson

47 Perception of affect in speech by older hearing aid users
Juliane Schmidt, Diana Herzog, Odette Scharenborg, and Esther Janse

Notes

49 Blank pages for notes

Index of Authors
Preface

Following the previous workshops on laughter held in Saarbruecken (2007), Berlin (2009), and Dublin (2012), we have the pleasure to welcome you to the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech in Enschede, the Netherlands on the 14th–15th of April 2015.

Non-verbal vocalisations in human-human and human-machine interactions play important roles in displaying social and affective behaviors and in controlling the flow of interaction. Laughter, sighs, filled pauses, and short utterances such as feedback responses are among some of the non-verbal vocalisations that have been studied previously from various research fields. However, much is still unknown about the interplay between phonetic and visual characteristics of non-verbal vocalisations, the relations to the intentions (encoding) and perceived meanings (decoding), and how non-verbal vocalisations can be automatically recognized and generated in human-machine interaction.

The goal of this workshop is to bring together scientists from diverse research areas and to provide an exchange forum for interdisciplinary discussions in order to gain a better understanding of laughter and other non-verbal vocalisations. The workshop consists of invited talks and oral presentations of ongoing research and discussion papers. We are very excited that Prof. dr. Anton Nijholt (University of Twente), Prof. dr. Catherine Pelachaud (CNRS, Telecom ParisTech), Dr. Disa Sauter (University of Amsterdam), and Dr. Jürgen Trouvain (Saarland University), four experts known for their excellent research, have accepted our invitations to speak about their research in relation to laughter. In addition, we have accepted 8 oral and 3 poster submissions. The invited keynote speakers and participants of the workshop have backgrounds in psychology, computer science, human-machine interaction, phonetics, and linguistics which we expect will give rise to lively cross-disciplinary discussions.

We would like to thank Charlotte Bijron, Alice Vissers-Schotmeijer, and Lynn Packwood (University of Twente) for their help in organizing this workshop. Our thanks also go the Human Media Interaction group of the University of Twente and the UT Aspasia Fund for their financial support.

Khiet Truong, Human Media Interaction, University of Twente, Enschede
Dirk Heylen, Human Media Interaction, University of Twente, Enschede
Jürgen Trouvain, Computational Linguistics and Phonetics, Saarland University, Saarbrücken
Nick Campbell, Speech Communication Lab, Trinity College Dublin, Dublin
Keynote

Prof. dr. Anton Nijholt – University of Twente

You had to be there to know why it was funny
Humor is important in our daily life, whether our activities are at home, at work, or in public spaces, for example during sports or other recreational and entertainment activities. Until now, computational humor, the research area that investigates rules and algorithms to understand and to generate humor, has hardly looked at other than verbal humor, in particular jokes. However, nowadays, humor has to be understood when it appears in digital audiovisual media, in interactive virtual environments (game environments), and in digital enhanced real-life environments. We explore whether future smart environments can be given a sense of humor. In order to do so we introduce some aspects of humor theory and how it can be employed in other than language dominated events. What can we learn from the employment of humor in comedy, movies and other ‘artificial’ worlds such as videogames and how can we use this knowledge in ‘semi-artificial’ worlds such as smart environments?

Bio Anton Nijholt received his PhD in Computer Science from the Vrije Universiteit of Amsterdam. He held various research positions in Canada, Belgium and the Netherlands before settling down as professor at the University of Twente in the Netherlands. He organized and acted as program chair of the main conferences on entertainment computing, affective computing and multimodal interaction. On these topics hundreds of publications, usually co-authored with his PhD students, have appeared. He is editor in chief of the journal Frontiers in Human-Media interaction and in recent years he has co-edited books on playful user interfaces, social collective intelligence and brain-computer interfaces. Also in recent years he started investigating how humor can profit from digital technology and how digital technology can profit from humor.
Keynote

Prof.dr. Catherine Pelachaud – CNRS Telecom ParisTech

LoL, Laughing with Greta

In this talk, I will first present our animation model of laughing for a 3D virtual agent. This model relies on data-driven approach; it simulates different types of movement: rhythmic and saccadic movement pattern; and it captures the dependencies between the movements across modalities. The laughing agent is used in LoL, a demo where a user interacts with a virtual agent able to copy and to adapt its laughing and expressive behaviors on-the-fly. In this demo, our aim is to study how copying strategies of the agent affects the user-agent interaction, his/her perception of the agent and his/her experience of the interaction.

Bio Catherine Pelachaud is a Director of Research at CNRS in the laboratory LTCI, TELECOM ParisTech. She received her PhD in Computer Graphics at the University of Pennsylvania, Philadelphia, USA in 1991. Her research interest includes embodied conversational agents, nonverbal communication (face, gaze, and gesture), expressive behaviors and socio-emotional agents. She has been involved and is still involved in several European projects related to multimodal communication (EAGLES, IST-ISLE), to believable embodied conversational agents (IST-MagiCster, ILHAIRE, VERVE, REVERIE), emotion (Humaine, CALLAS, SEMAINE, TARDIS) and social behaviors (SSPNet). She is member of the Humaine Association committee. She is associate editors of several journals among which IEEE Transactions on Affective Computing, ACM Transactions on Interactive Intelligent Systems and Journal on Multi-modal User Interfaces. She has co-edited several books on virtual agents and emotion-oriented systems.
Keynote

Dr. Disa Sauter – University of Amsterdam

Nonverbal vocalisations are basic human expressions of emotion

Although great advances have been made over the last decades in the scientific understanding of emotional communication, research has been strongly biased towards the investigation of visual signals, especially facial expressions. In this talk, I present a set of studies examining vocal expressions of emotions, specifically, nonverbal vocalisations, such as screams and laughs. My findings show that these signals reliably communicate a range of emotional states via distinct constellations of acoustic cues. Furthermore, I will demonstrate that some of these vocalisations constitute universal signals of emotions, and that the mappings between some emotions and vocalisations are innate. In light of these data, I will propose that some emotional vocalizations constitute basic human expressions of emotion, including of several distinct positive emotional states.

Bio Disa Sauter is an Assistant Professor in the Department of Social Psychology at the University of Amsterdam. Disa did her BSc in Psychology and Cognitive Science at University College London (2002), followed by a PhD in the same department (2006). She then worked as a post-doctoral researcher at the Institute of Psychiatry, King’s College London, after which she received an ESRC fellowship, which she took up at Birkbeck College London. In 2008, Disa moved to the Netherlands to work as a staff researcher at the Max Planck Institute for Psycholinguistics. Since 2011 she is at the University of Amsterdam, where she holds a Veni fellowship from the Dutch Science Foundation.
Keynote

Dr. Jürgen Trouvain – Saarland University

Research on laughter in vocal communication and its progress in the last decade

The beginning of this workshop series goes back 10 years from now. Since then the research on laughter as a vocal signal has been made a substantial progress and is reflected by numerous studies by colleagues from different labs presented at various workshops, conferences and journals. This personal re-view attempts to enlighten the scientific achievements of the last decade on a topic largely ignored before that period. The talk will also include some ideas on research questions which still seems to be underexplored.

Bio After his PhD 2003 in Phonetics from Saarland University, Saarbrcken, Germany, with a thesis on tempo variation in speech production he worked in past and present as a lecturer and researcher in various projects on speech synthesis and language learning at Saarland University and at the University of Luxembourg. Currently he leads a French-German project on individual feedback in computer-assisted spoken language learning. His main research interests include non-native speech and phonetic learner corpora, stylistic and dialectal variation of speech, historical aspects of speech communication research, and nonverbal vocalisations.
AUDIO-VISUAL LAUGHTER SYNTHESIS SYSTEM

Hüseyin Çakmak, Kevin El Haddad, Thierry Dutoit

University of Mons (UMONS/Belgium)
{huseyin.cakmak},{kevin.elhaddad},{thierry.dutoit}@umons.ac.be

ABSTRACT

In this paper we propose an overview of a project aiming at building an audio-visual laughter synthesis system. The same approach is followed for acoustic and visual synthesis. First a database has been built to have synchronous audio and 3D visual landmarks tracking data. Then this data has been used to build HMM models of acoustic laughter and visual laughter separately. Visual laughter modeling was further separated into a facial modeling and head motion modeling. An automatic laughter segmentation process has been used to annotate visual laughter. Finally, simple rules were defined to synchronize all the different modalities to be able to produce new durations.

Keywords: audio, visual, laughter, synthesis, HMM-based

1. INTRODUCTION

Among features of human interactions, laughter is one of the most significant. It is a way to express our emotions and may even be an answer in some interactions. In the last decades, with the development of human-machine interactions and various progress in speech processing, laughter became a signal that machines should be able to detect, analyze and produce. This work focuses on laughter production and more specifically on the synchronization between audio and synthesized visual laughter.

As summarized in Figure 1 the main of this project is to build a complete audio-visual laughter synthesis system. From an audio-visual database, HMM-based modeling is done for audio data, face and head motion separately with their respective annotations. Phonetic transcriptions were made manually while face transcriptions are based on a GMM-based clustering of the visual data. Head transcriptions are derived from face transcriptions and from head motion data automatically as well.

Figure 1: Overview of the pipeline for HMM-based audio-visual laughter synthesis

The paper is organized as follows: Section 2 gives a brief overview on the database used in the project, Section 3 explains the acoustic laughter synthesis method, Section 4 explains the visual synthesis, Section 5 describes synchronization rules between audio and visual modalities and Section 6 concludes and gives an overview of current and future work.

2. THE AV-LASYN DATABASE

The AVLASYN Database [1] used in this work is a synchronous audio-visual laughter database designed for laughter synthesis. The corpus contains data from one male subject and consists of 251 laughter utterances. Professional audio equipment (Shure SM58 micro and RME Fireface 400 Soundcard) and a marker-based motion capture system have been used for audio and facial expression recordings respectively. Figure 2 gives an overview of the recording pipeline.

The database contains laughter-segmented audio files in WAV format and corresponding motion data in the Biovision Hierarchy (BVH) format. A visual segmentation was done on laughter files from which audible parts were phonetically annotated. The laughs were triggered by watching videos found on the web. The subject was free to watch whatever he wanted. A total amount of 125 minutes were watched by the subject to build this corpus.
This led to roughly 48 minutes of visual laughter and 13 minutes of audible laughter. Audible laughter is less than visual laughter because smile is also part of visual laughter and has no audible components. Also, facial expression begins before acoustic expression of laughter and continues after the last audible sound.

3. ACOUSTIC LAUGHTER SYNTHESIS

Acoustic synthesis of laughter using Hidden Markov Models (HMMs) has been addressed in 2013 [11]. To characterize the acoustic laughter, phonetic transcriptions [12] were used and the results outperformed the state of the art. Extensions of the latter work were done to perform automatic phonetic transcriptions [10] and to integrate the arousal in the system [9]. The goal of audio laughter synthesis is to generate an audio waveform of laughter.

Several versions of HMMs were developed, with varying contextual information and algorithms for estimating the parameters of the source-filter synthesis model. These methods were compared, in a perception test, to human laughs and copy-synthesis laughs. In this evaluation, participants were invited to rate the naturalness of the laughs they were listening to. The evaluation showed that 1) the addition of contextual information does not increase the naturalness, 2) the proposed method is significantly less natural than human and copy-synthesized laughs, but 3) significantly improves laughter synthesis naturalness compared to the state of the art. The evaluation also demonstrates that the duration of the laughter units can be efficiently learnt by the HMM-based parametric synthesis methods.

4. VISUAL LAUGHTER SYNTHESIS

In [4] we have proposed a visual laughter synthesis system. This work has shown that a separate segmentation of the laughter is needed to correctly model the visual trajectories meaning that phonetic transcriptions are not suited to describe the visual cues for laughter as it has been shown to be feasible for speech [5, 7, 6, 8]. Further developments have shown that the head motion should be modeled separately as well [2]. The general visual laughter synthesis pipeline is given in Figure 3. Principal Component Analysis (PCA) is performed independently on face and head motion data to uncorrelate the features and to reduce dimensions. Then the reduced data are trained separately for head and face with their specific transcriptions. As shown on Figure 4, the phonetic transcription are more granular than visual transcriptions. Facial transcription are mainly limited to two classes: neutral or laughter. Head motion transcription are based on the facial transcription in the sense that the neutral remains the same and the laughter is subdivided into a set of identical classes which represent each one period of the head oscillation occurring during laughter.

5. SYNCHRONIZATION RULES

In [4], the synchronization between modalities was guaranteed by imposing synthesized durations to be the same as in the database, in which the transcriptions are synchronous in the first place. To bring this to the next level and to be able to synthesize audio-visual laughter with any wanted duration, we derived simple synchronization rules to model the relationships between transcriptions as explained below.
As explained in the previous sections, audio, facial data and head data are modeled separately with their own transcriptions and this leads to the need of synchronization techniques. The basic principle lying under the proposed method in [3] is the study of the relation between the audio and visual transcriptions. Rules are extracted from the study of temporal shift between the beginning of the perceptible visual laughter and the beginning of the audible laughter. Likewise, rules are extracted from the study of temporal shift between the end of visually perceptible laughter and the end of the post-laughter inhalation. This method makes it possible to generate visual transcriptions starting from audio transcriptions. An online MOS test is then conducted to rate the quality of the animation and the matching between audio and visual modalities. The results show that there is no significant difference between original animations and those using the synchronization rules of this section.

6. CONCLUSION AND FUTURE WORKS

This paper briefly presented recent advances towards building an audio-visual laughter synthesis system. The recording of a database, the development of the acoustic and visual parts of the system has been introduced. Synchronization rules used to be able to unify audio and visual parts has been presented as well.

Future works include the integration of the arousal as a control of the intensity of the laughter to be produced. The extension to more complex laughs is also planned, both in terms of synthesis and synchronization. Building the visual transcription directly from the audio track of laughter rather than being dependent on the availability of its phonetic transcription would be valuable as well.

7. ACKNOWLEDGEMENT

H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l’Industrie et l’Agriculture (F.R.I.A.), Belgium.

8. REFERENCES

TOWARDS A SPEECH SYNTHESIS SYSTEM WITH CONTROLLABLE AMUSEMENT LEVELS

Kevin El Haddad, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit

TCTS lab-University of Mons (UMONS/Belgium)
{kevin.elhaddad},{huseyin.cakmak},{stephane.dupont},{thierry.dutoit}@umons.ac.be

ABSTRACT

This work exposes some observations on "shaking vowels". These vowels containing a tremolo-like sound are present in amused speech. Some of the temporal and spectral patterns are exposed and observations are given concerning them. This work is a preliminary step towards being able to use these "shaking vowels" along with speech-smiles, speech-laughs and laughter to control the degree of amusement in synthesized speech.

Keywords: Laugh, Amused speech synthesis, smile-laugh continuum

1. INTRODUCTION

Amusement can be expressed vocally and through facial expressions. The most common visual amusement expression is smiling. Vocally, amusement is commonly expressed with laughter but also through smiles. Indeed, when occurring along with speech, smiles are audibly recognizable as shown in [1] and [11]. The speech style created from speech modulated by smiles is called speech-smiles. Laughter often occur as isolated events, but bursts of laughter can also be intermingled with speech. In the latter case, the uttered word or sentence is referred to as speech-laugh. The subject of this paper concerns amusement expressed at the same time as speech. An important hypothesis to take into account here is the existence of a smile-laugh continuum. In [12], based on an analysis of facial expression, Ruch proved that Duchenne smiles (smiles related to enjoyment) were involved in laughter. The smile-laughter continuum hypothesis could be viewed as an amusement level scale axis. This axis would have neutral speech on one end and high intensity isolated laughter on the other. It would also contain consecutively speech-smiles and speech-laughs in its increasing direction.

Several studies in different areas have been conducted concerning speech-smiles and speech-laughs. The lip deformation effects on speech have been studied in [5]. In [7], Lasarcyk studied the cues contributing to the vocal discrimination of smiles. The pitch has been found to have the highest impact on people recognition of smiles in [7]. In [1] a perceptual study was conducted to evaluate the abilities of listeners to recognize different kinds of smiles. Speech-smile synthesis systems have also been developed using Hidden Markov Models in [3] and [4]. Speech-laughs acoustic characteristics have been analyzed in [12], [6], [8] and [2]. These studied the formant frequencies, pitch variations and durations of speech sounds when laughter occurs. Attempts for speech-laugh synthesis can also be found in [9] and [4].

We report some early data exploratory results from a database we collected and which has for purpose, among others, to improve the amused speech synthesis systems developed in [4] and [3]. The purpose of the recordings is the collection of speech-smiles but, sentences containing vowels altered by some kind of tremolo were also collected. The purpose of this study is to provide some of the foundations for future work on controlling the degree of amusement in synthesized speech. If speech-laughs are considered to exist at different levels of amusement arousal, this type of sentences would be considered as one of the degrees of amusement in speech. The purpose of this study is to provide one of the foundations for future work on controlling the degree of amusement in synthesized speech.

2. DATA DESCRIPTION

In order to record a speech-smile database, an actor was asked to sound amused (but without laughing) while reading sentences. The actor, probably focused too much on being in an amused state, sometimes uttered speech-laughs. In some of those sentences, the vowels were altered by some sort of tremolo. As will be seen later, these vowels will be considered in future studies as laughter bursts (by laughter bursts we refer to the definition given by Ruch [10]) with no or very short syllabic pulsations. For this reason, these altered vowels will be referred to as "shaking vowels" to differentiate them from the ones in the speech-laughs containing syllabic pulsa-
Figure 1: Temporal and spectral representations of three shaking vowels. (a) represents e, (b) represents ˜a and (c) represents ɛ. In the upper and lower parts you can see the temporal representations of the waveform and the pitch respectively. In the lower part you can also see the shaking vowels’ spectrograms.

Given the original goal of the database recording, whenever such shaking vowels were noticed, the actor was asked to repeat the same sentence again until a smiled sentence containing no such tremolo was obtained. In the end though, among the 105 non speech-smile sentences, 43 were counted as containing shaking vowels, leading to a sub-database used in this study. We make some examples of those, along with the corresponding speech-smile sentences, here http://tcts.fpms.ac.be/~laughter/laughterWorkshop15.

The shaking vowels found in that sub-database were present in the French vowels/sounds a, e, i, o, ɛ, u, y and ˜a.

3. RECORDING DETAILS

The speech-smiles recordings took place in a sound proof room using a Rode Podcaster microphone. They were made at a sampling frequency of 48kHz and encoded in a PCM 16 bits WAV format. The actor was asked to read sentences in French.

4. OBSERVATIONS AND FURTHER USE

In this section, we detail some of the shaking vowels observed characteristics.

Fig. 1 shows the most commonly found pattern characterizing the shaking vowels. The upper part of Fig. 1 (a), (b) and (c) shows the waveform of the shaking vowels while the lower part shows their spectrogram and a pitch estimation curve in red. We can see the same pattern for three different vowels (a, ɛ and ˜a). By listening to these shaking vowels, we concluded that this pattern was formed so to say by the "concatenation" of two instances of the same vowel. Alternatively this could be described as a modulation of the vowel normally stable portion. Other patterns were also found, for example in Fig. 2 (for the french vowel a) where more than two instances of the same vowel are seen to be "concatenated".

Some observations can be made about this pattern:

1. one of the concatenated vowels has a higher amplitude than the other. In most cases in our study, the second concatenated vowel’s amplitude was more attenuated than the first. But some other cases were found where it was the first one’s amplitude that was more attenuated than the second as shown in Fig. 3.

2. the distance between those concatenated vowels vary from one shaking vowel to another as can be seen when comparing Fig. 1. In Fig. 1-(a), the vowels are close to each other and the transition between them is continuous as can be seen in the corresponding spectral
representation. In Fig. 1-(b) and Fig. 1-(c) the vowels are further apart (slower modulation) from each other than in Fig. 1-(a) but the transition between them is still continuous.

No conclusion could be drawn concerning the pitch. Further studies should be made to fully understand its behavior in shaking vowels. Those observations will be taken into account for further work on amused speech synthesis. Indeed, the depth of the amplitude excursions in a shaking vowel could be an interesting parameter to control. The time difference between the peak amplitude excursions in shaking vowels is another interesting parameter to exploit. Further studies could also assess whether or not those amplitude differences and timings of these modulations correlate with the degree of amusement. If so, a pulse of air could eventually be inserted in between vowels amplitude peaks. This threshold would mark a transition between two levels of amusement in speech-laughs. In other words and as mentioned earlier, the shaking vowels will be considered as laughter bursts with no or very short syllabic pulsations.

Figure 3: Shaking vowel pattern in which the first vowel has a lower amplitude than the second.

5. CONCLUSION

In this study, we exposed some observations about shaking vowels. Their use for the synthesis of more expressive amused speech was also proposed. Two features were found to be interesting for exploitation in future work, the depth of vowels amplitude excursions, as well as their timing. Future work will include those shaking vowels in a synthesis system with a further goal to control the degree of amusement in speech.

6. REFERENCES

USING MAGE FOR REAL TIME SPEECH-LAUGH SYNTHESIS

Kevin El Haddad, Alexis Moinet, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit

TCTS lab-University of Mons (UMONS/Belgium)
{kevin.elhaddad},{alexis.moinet},{huseyin.cakmak},{stephane.dupont},{thierry.dutoit}@umons.ac.be

ABSTRACT

In this paper, we present an ongoing work which aims at synthesizing speech-laugh sentences in real-time. To do so, the Hidden Markov Model (HMM)-based speech-laugh synthesis system will be used along with the MAGE software library. First results are available online on tcts.fpms.ac.be/~laughter/laughterWorkshop15.

Keywords: Real-time speech-synthesis, speech-laughs, laugh

1. INTRODUCTION

Recent research has been focusing on improvements to the expressivity and naturalness of synthetic speech. One way to do so is to add emotions to the synthesis. This paper presents work in the framework of affective speech synthesis, focused more particularly on amused speech. We attempt to present perspectives on real-time speech-laugh synthesis. Speech-laugh designates laughs happening at the same time as speech and intermingled with it. Several phonetic and acoustic studies concerning speech-laugh can be found in [11], [7] and [5]. To the best of our knowledge, very few were made addressing speech-laugh synthesis. Lasarcyk [6] reproduced natural speech-laughs using an articulatory synthesizer. Also, this work relies on a previously developed Hidden Markov Model (HMM)-based synthesis system for speech-laugh [4]. We will attempt to use this system alongside with the MAGE software [1] to synthesize and control speech-laugh in real-time. This system will be integrated in the Chist-ERA Joker project. This project is currently designing JOKER, a generic intelligent user interface providing a multimodal dialogue system with social communication skills including humor, empathy, compassion, charm, and other informal socially-oriented behaviors (more information about this project can be found in http://www.chistera.eu/projects/joker).

This paper presents the HMM-based speech-laugh synthesis system in section II. We then present the MAGE software library in section III. In section IV, we explain how the two systems will be used together. Finally, section V will conclude.

2. HMM-BASED SPEECH-LAUGH SYNTHESIS SYSTEM

This section only summarizes the system presented in [4]. Please refer to that earlier work for a more detailed presentation. This work attempts to model speech-laughs. Speech-laughs are variable and depend on social and/or situational context. A realistic model would be very complex to create. So further studies, like the one in chapter 9 of [6], in order to develop such a model. Our first approach was therefore, to simplify the model by regarding speech-laughs as laughter bursts replacing vowels in speech-smiles. Fig. 1 shows the HMM-based speech-laugh synthesis system workflow.

![Figure 1: HMM-based speech-laugh synthesis system](image-url)
Likelihood Linear Regression (CMLLR) algorithm [2]. Moreover speech-smile HMM models were also trained from a smaller corpus (approximately 12 minutes long) of speech-smiles recorded from another French-speaking person. This person was trained to pronounce the desired smiled voice before the recordings. The synthesis results in [4] and in [3] showed the efficiency of those data when used for training HMM models of speech-smiles.

Vowels are then replaced by laughter bursts inside those speech-smile sentences. In order to do that, "laughing vowels" which are laughter bursts occurring inside vowels, were recorded from the same person the speech-smile sentences were recorded from. He was asked to produce sustained French vowels while watching funny videos. HMM models are then also created for the laughing vowels. These models are used to replace the vowels by laughing vowels in the synthesized speech-smile sentences. The sentences synthesized using this system proved to be efficiently perceived as amused as the evaluations in [4] show.

Practically, for synthesis, the input of the system is a list of phonemes. The system then chooses the best suited trajectories for the features (that model the voices) using a maximum-likelihood parameter generation algorithm making use of the HMM model previously made during training [10]. From those trajectories, a waveform is generated from a synthesizer. Fig. 2 shows the laughing vowels labeling.

3. MAGE

MAGE [1] is a software library built as a realtime layer around HTS engine [9]. It allows modifications of the parameters and inputs of an HMM-based speech synthesizer on-the-fly, therefore affecting its output while it is speaking.

For instance, a user can give, one-by-one, a sequence of phonemes (complying with the format specified in [8]) as inputs of the system and it will start talking as soon as it receives the first phonemes. Then the user can change the pitch trajectories, or the vocal tract length of the voice as it is produced. When no more phonemes are sent to MAGE, it can either sustain indefinitely the last received one or just stop talking.

4. REAL TIME SPEECH-LAUGH SYNTHESIS

Since phonemes can be processed by MAGE progressively, whenever we want to add a laughter in a sentence, we intend to replace a given vowel with a sequence “pulse-vowel-pulse-vowel” just before it is sent as an input. The readers are invited to download some first examples synthesized using MAGE while this paper was being written on tcts.fpms.ac.be/~laughter/laughterWorkshop15

5. CONCLUSION

In this extended abstract, we presented our ongoing work concerning real-time speech laugh synthesis. A concrete example of application will be its integration into the Chist-ERA Joker Project cited above.
6. REFERENCES


EXPLORING THE BODY AND HEAD KINEMATICS OF LAUGHTER, Filled pauses and breaths
Spyros Kousidis, Julian Hough and David Schlangen
Dialogue Systems Group, Bielefeld University, Germany
spyros.kousidis@uni-bielefeld.de

ABSTRACT
We present ongoing work in the DUEL project, which focuses on the study of disfluencies, exclama-
tions, and laughter in dialogue. Here we focus on the multimodal aspects of disfluent vocalizations, namely laughter and laughed speech, filled pauses, and breathing noises. We exemplify these phenomena in the rich multimodal Dream Apartment Corpus, a natural dialogue corpus, which, in addition to comprehensive disfluency and laughter annotation, comprises tracking data for the body and head. We discuss possible directions for developing models that can perceive as well as generate such multimodal behaviour.

Keywords: Laughter, disfluencies, multimodal

1. INTRODUCTION
In interaction research on both humans and artificial agents, there is a growing interest in phenomena that were previously avoided or ignored, namely disfluencies and laughter. The motivation comes both from the need for more precise perception of user behaviour, as well as the goal of modeling natural interaction and designing agents capable of it [15].

There are a number of improvements that modelling disfluencies and laughter can afford to human-agent interaction. For example, filled pauses at the beginning of turns may act as a signal that more time is needed for production [2]. [1] demonstrated how this function of filled pauses could be exploited in human-robot interaction. Similar interaction management functions of audible breathing at turn boundaries can also be exploited. [7] show how the next speaker in multiparty interactions can be predicted using breathing and gaze features.

Laughed speech is problematic for ASR [3] and thus detection of laughed speech may help deal with possibly erroneous ASR output. [13] presented a classification of laughter, filled pauses, speech and silence, using audio features and a bigram model. Detection of filled pauses and repair disfluencies, which aids parsing spontaneous speech, has recently focussed on word-by-word incrementality [6], in order to allow the online interpretation of user speech beyond idealised, fluent utterances.

The multimodal aspects of disfluency and laughter phenomena have also had recent interest: for example laughter detection from facial features has proven quite successful [14, 4]. However, it is interesting to explore the behaviour in other modalities during such episodes. [10] found that human raters were highly confident in distinguishing laughter from non-laughter, when observing a 3D virtual puppet performing body animations derived from raw human motion capture data. [5] classified different types of laughter (labeled by naive annotators on 3D animated avatar clips) using only the motion capture data. In both cases, the avatars were very simple, so that only the movements of the body were visible, while any other features such as body shape, height and face had been removed. The animations were based on episodes of elicited laughter, rather than laughter that occurs in interactive settings.

Using a multimodal corpus of natural dyadic interactions, we employ analytical methods to address the challenges outlined above, with the general hypothesis that a combined approach to detecting disfluencies and laughter phenomena, using multimodal data, will fare better than the previous approaches detecting these phenomena individually. As part of this ongoing work, we present our initial exploration in the area of laughter, filler and meaningful breath detection using kinematic features of the head and body.

2. MATERIAL AND ANNOTATIONS
In this work, we use the Dream Apartment Corpus [8, 9], a corpus of dyadic interactions between German speakers that features audio, video, body, head, eye and gaze tracking in a “minimally invasive” setting (no worn sensors, except for head-worn microphones). Body tracking was performed with Microsoft Kinect for Windows (version 1) and head.eye tracking was performed with Seeingmachines FaceLab. An example scene from the corpus is shown in Figure 1. The Kinect data comprises 20 joints per
body at a sampling rate of 30 Hz, while the FaceLab data has twice that sampling rate (60 Hz).

The task of each dyad is to discuss the layout of an apartment in which they would co-habit, given an extraordinary amount of funds. This task is in place mostly to induce spontaneity in the dialogue by giving a topic and a limited amount of time (15 minutes) to come up with a proper apartment plan. The entire corpus consists of 9 dyadic interactions (balanced genders, varying degrees of acquaintance) approximately 20 minutes each (3 hours in total).

Based on the audio, we have annotated all instances of disfluent and laughter phenomena in the corpus as shown in Table 1. Audible breath and laughter offset, a special type of loud inhalation that often (but not always) follows laughter episodes, as well as laughed speech and laughter are annotated using XML-like tags. Filled pauses are enclosed in curly brackets and also start with an F character accompanied by the filled pause transcription.

Laughter is accompanied by movements of the body and head, which motivates efforts to use this modality in order to identify laughter events in dialogues. [11] have proposed the use of a Body Laughter Index which uses features such as the kinetic energy of the head and shoulders, as well as the periodicity of the movements of these joints. We are interested in improving on such metrics using the available motion capture data in the DAP corpus. Figure 2a shows the kinetic energy per joint during a laughter episode. The kinetic energy of each joint is computed separately from a vector of displacement per unit time derived from the raw Kinect tracking data.

A second area of interest is the periodicity of body movements during laughter episodes [12]. A second laughter episode is shown in Figure 2b. We observe the synchronous oscillation of the shoulders, neck and head, at a frequency of 8 Hz which is higher than the approximate 5 reported commonly in the literature. Of further interest is the study of periodicity features (frequency and phase) of laughter pulses and their comparison to periodicity features of the body and head.

4. HEAD GESTURES

Laughter and/or laughed speech is often accompanied by head gestures. The DAP corpus has already been annotated for communicative head gestures [8], by two annotators per dialogue. We consider only the overlapping annotated segments for which there is agreement on the type of gesture performed among the annotators, and compare these with carefully segmented occurrences of laughter, filled pauses, and breaths. The results are shown in Table 2. We find that laughter episodes are accompanied by head gestures 39% of the time, which is quite frequent. The gesture types vary (the annotations in [8] distinguish 9 different types) and further work is required in order to determine whether some

<table>
<thead>
<tr>
<th>phenomena</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laughter</td>
<td>&lt;Laughter/&gt;</td>
</tr>
<tr>
<td>Laughed Speech</td>
<td>&lt;Laughter&gt; speech &lt;/Laughter&gt;</td>
</tr>
<tr>
<td>Filled Pause</td>
<td>{F ähm } {F äh}</td>
</tr>
<tr>
<td>Breath</td>
<td>&lt;Breath/&gt;</td>
</tr>
<tr>
<td>Laughter Offset</td>
<td>&lt;LaughterOffset/&gt;</td>
</tr>
</tbody>
</table>

Table 1: Subset of Annotation Labels from the DUEL annotation manual.
gestures accompany laughter more frequently than others. It is clear, however, that any automatic laughter detection that uses head motion features must take into account simultaneous communicative head gestures and their influence in head kinematics.

Similarly, although less often, we observe co-occurrence of breaths, laughter offsets and filled pauses with head gestures. The relative frequency of head turns as opposed to other gestures is higher in comparison to laughter episodes. We attribute this to instances of one interlocutor looking away from the other while audibly breathing or uttering a filler, in order to gain time to think. This analysis can be made more precise by considering events at the end of vocalisations, and also looking at the gaze vector, in addition to the head movement.

5. OUTLOOK

Our initial explorations in a rich multimodal corpus indicate several potential contributions and improvements on the current state of the art in the area of body-motion-based laughter and disfluency detection. As part of the DUEL project, data collections similar to the DAP corpus in three languages (German, French, Chinese) are planned, in order to explore cross-linguistic aspects of disfluencies and laughter in elicited spontaneous dialogues.

6. REFERENCES


---

Table 2: Frequency of simultaneous head gesture and laughter/laughed speech, fillers and breaths per gesture type.

<table>
<thead>
<tr>
<th>Gesture</th>
<th>Laughter</th>
<th>Fillers</th>
<th>Breaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nod</td>
<td>20</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Turn</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Tilt</td>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Shake</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Protrusion</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Retraction</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Shift</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Bobble</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Jerk</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Co-occur % 39.2 14.6 12.4

---

1 http://www.eyetracking.com/Hardware/ Eye-Tracker-List
THE RELATIONSHIP BETWEEN LAUGHTER INTENSITY AND PERCEIVED HUMOUR

Gary McKeown, Will Curran
School of Psychology, Queen’s University Belfast
g.mckeown@qub.ac.uk, w.curran@qub.ac.uk

ABSTRACT

There is a long history of recognizing a continuum in social signals of positive affect, with the continuum ranging from mild amusement signals to strong laughter. However there has been little systematic effort to assess what this continuum might mean. We present data that shows that incorporating intensity measures of laughter into laughter research is an important component and that there is a strong relationship between laughter intensity and humour. This may be intuitively obvious but the strength of the relationship suggests that intensity measures should be included in all laughter research.

Keywords: Laugh Intensity; Humour; Social Signals

1. INTRODUCTION

“A graduated series can be followed from violent to moderate laughter, to a broad smile, to a gentle smile, and to the expression of mere cheerfulness.” Darwin [1] noted that there was something that resembled a continuum stretching from intense laughter to milder cheerfulness expressions. This point has also been made in the context of comparative psychology [2]. What tends to distinguish the ends of this continuum is the level of intensity that is associated with the signals. We intuitively have a feel for what is meant by intense laughter, but very little research has looked at it in detail and what it might mean or has questioned the assumptions. It is such an important component of laughter as a social signal that it is imperative to have clarity on what intense laughter might mean. In this paper we report on research conducted as part of the ILHAIRE project, which examined laughter intensity and what it might mean. We also offer some options for operationalizing and measuring the intensity of laughter as a construct for research.

2. WHAT IS LAUGHTER INTENSITY

A number of ways have been proposed of operationalizing laughter intensity. Lynch [3] used the facial action coding system [4] intensity rating to provide a measure of intensity; Hess, Banse and Kappas [5] used facial electromyography and Urbain et al. [6] synthesized laugh intensity based on a variety of measures of emotional arousal. In the study reported in this paper we decided not to impose too many assumptions as to what features comprise laughter intensity. Instead as an initial assessment we chose to simply assume that humans—as long term laughers and receivers of laughter as a social signal—can be deemed to have developed some expertise in this domain throughout the course of their lives. In this regard, we asked a large number of human participants to rate the intensity of a given laugh presented as an audio-visual clip. This has the added benefit of allowing us to ask other questions regarding the nature of a laugh as deemed by our “laugh experts”, allowing us to compare ratings of various features.

2. LAUGHTER STIMULI AND RATINGS

Laughs were collected from the “Belfast Storytelling Database” a corpus of groups of three or four participants being recorded audio-visually while sitting around a table telling each other stories [7]. There were six conversational sessions, each lasting approximately 1 hour, and 21 participants. Three of the sessions were conducted in English by native English speakers and three were conducted in Spanish by native Spanish speakers. Visual recordings were taken from a camera positioned in front of each individual, capturing the face and upper torso. Acoustic recordings were captured from head mounted microphones. These signals were combined to make audiovisual recordings of each participant. A series of segments of laughter were extracted from the individual recording of the group conversational interactions. The goal in extracting these laugh segments was to remove as much of the conversational contextual information as possible. For this study the number of segmented laugh stimuli was 310 for the English speaking laughs drawn from 9 participants and 255 for the Spanish speaking laughs drawn from 12 participants.

The individual laughs were then placed on a website and rated along a series of dimensions. These dimensions asked a number of questions: 1. Can you rate the intensity of the laugh on a 10 point scale, from 1 no intensity to 10 maximum intensity?
2 & 3. Do you think the laugh was more related to something humorous or more the kind of laugh that would occur in normal conversation? Can you indicate how much by marking along the 10 point scale? 1 means not related to humorous or conversational and 10 means completely related to humorous or conversational. Place mark in both scales please.

4. How malicious do you think the laugh was on a scale of 1 to 10? 1 means not malicious at all and 10 means extremely malicious.

5. How benevolent do you think the laugh was on a scale of 1 to 10? 1 being not benevolent at all and 10 being extremely benevolent.

laughs were judged by raters recruited through Amazon’s Mechanical Turk website. In the Spanish laugh sample 147 unique raters provided 1902 ratings for each of the five dimensions. In the Spanish laugh sample 540 unique raters provided 5690 ratings for each of the five dimensions.

3. RESULTS

As intensity is the most important dimension for the purposes of this paper, analysis is based on intensity and its relationship with the other dimensions. Figures 1 and 2 show the frequency distributions of the intensity ratings for English and Spanish laughter. These highlight that the modal intensity rating is approximately three with a steep drop off for high intensity ratings.

Table 1 presents a correlation matrix of each of the variables (Pearson’s r). English correlations are presented in the upper triangle and Spanish in the lower triangle.

All correlations are significant at the p<.01 level except for Malicious and Benevolent where there is a zero correlation. That these are almost all significant is unsurprising given the sample sizes. What is more interesting is the effect sizes (r is itself an effect size measure). We see interesting consistency in effect size measures across the two samples between Intensity and Humourous, r=0.7 (English) and r=0.67 (Spanish), which is a very strong effect size; and between Intensity and Benevolence, r=0.29 (English) and r=0.3 (Spanish), which is a medium effect size. Figure 3 and Figure 4 present the Intensity-Humorous relationships as scatterplots with the points randomly jittered to more clearly display the distributions.
These results seem to present, perhaps obviously, a very strong positive relationship between a laugh’s rated intensity and the degree to which it is associated with humour. There is also a medium effect size relationship between laugh intensity and benevolence but, as we can, see humorous and benevolence are also correlated. To assess whether the humorous and benevolence variables are tapping into the same construct or whether benevolence adds some additional explanation beyond the humorous variable we compare two regression models for each sample. A simple model with intensity as a response variable and humorous as a predictor, and a more complex model that includes both humorous and benevolence as predictors. For both the English and Spanish samples there are small but significant contributions to the ratings of intensity by the benevolence variable beyond those provided by the humorous variable.

Table 2: Comparison of simple and complex regression models for English and Spanish samples.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple</td>
<td>Complex</td>
</tr>
<tr>
<td>Humour</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>SE</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Benevolent</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>SE</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$R^2_{adj}$</td>
<td>0.48</td>
<td>0.49</td>
</tr>
</tbody>
</table>

As the $R^2_{adj}$ figures in Table 2 show, the humorousness ratings explain about 48% of the variance in intensity ratings in the English sample and 45% of the variance in intensity ratings in the Spanish sample with an additional 1% of the variance being explained by the benevolence variable in both samples. Thus there is a small contribution of the benevolence variable beyond that explained by the humorousness variable, but perhaps not enough to warrant inclusion.

4. DISCUSSION

In this paper we have argued that intensity is an important variable that should be considered in the research of laughter. Within two samples in the current data it appears to strongly distinguish the degree to which laughter is associated with humour, with high intensity laughs strongly associated with humour and low intensity laughs only weakly or not at all associated with humour. These low-intensity laughs are most likely the laughs that have mostly conversational functions [8] however confirmation of this is beyond the scope of the present paper. Given the strength of the relationship between intensity and humour ratings in this and other similar studies [9] we recommend that some measure of intensity is always included in studies that seek to investigate laughter.

5. REFERENCES

WHEN DO CHILDREN SMILE? PREDICTING GENUINE SMILES AND NON-GENUINE SMILES AMONG WINNING AND LOSING CHILDREN

Phoebe Mui, Martijn Goudbeek, Marc Swerts
Tilburg University
{h.c.mui, m.b.goudbeek, m.g.j.swerts} @ uvt.nl

ABSTRACT

In a production experiment, we elicited and coded facial expressions of Dutch and Chinese children who won and lost in a game. Results showed that social context predicted occurrence of genuine smiles: Children who played in pairs (versus alone) and children who won (versus lost) were more likely to show genuine smiles. Moreover, occurrence of non-genuine smiles was predicted by an interaction between social context and game outcome. Children who lost in pairs were mostly likely to produce non-genuine smiles; winning or losing could not predict non-genuine smiles among children playing alone. Cultural background of children (Dutch or Chinese) could predict neither the occurrence of genuine smiles nor that of non-genuine smiles.

Keywords: smile, children, social context, culture.

1. INTRODUCTION

Smiling is subject to contextual factors, such as the sociality of the situation. Presence of an audience, either real or imagined, has been found to augment smiling [4]. The effect of sociality on smiles has been observed in competitive events such as sports as well. Winning athletes smile more during social interaction than in the absence of social interaction (cf. [3], [7] and [8]). Such smiles, which accompany positive affect, are often considered genuine or Duchenne smiles [3].

The fact that sociality enhances smiling in winning situations raises the question as to whether the same effect can be observed for displays of defeat. Non-Duchenne smiles, also known as non-genuine or masking smiles [2], are often displayed when negative affect is actually being experienced, or in situations that call for appeasement. However, no unique display of defeat seems to exist; it consists of sadness, nothing, or contempt [8]. Thus, it remains unclear as to whether non-Duchenne smiles are defining features of expressions of losing.

It has been suggested that the Duchenne marker is not a universal marker of smile authenticity. In [10], French-Canadians were reported to rely more on the Duchenne marker in judging the authenticity of smiles than did mainland Chinese and Gabonese. These findings suggest possible cultural differences regarding the production and perception of smiles. Indeed, cultural display rules, which are societal expectations governing what expressive behaviour is appropriate, may play a role in smile production. In collectivist cultures (such as China; individualism score = 20; cf. [6]), display rules dictate that negative emotions be masked as they might impede group harmony. In cultures that are more individualistic (such as the Netherlands; individualism score = 80), the articulation of negative emotions is less frowned upon. In addition, the distinction between masculine and feminine cultures (cf. [6]) also lends itself to cultural differences in expressivity norms. Masculine cultures such as China (masculinity score = 66) prize success to a greater extent than feminine cultures such as the Netherlands (masculinity score = 14; cf. [6]). As a result, expressions of defeat are less encouraged in masculine cultures.

Taken together, it is reasonable to expect displays of defeat to be characterised by non-genuine smiles in cultures that are collectivistic or masculine. On the contrary, in cultures that are individualistic or feminine, expressions of negativity and failure are more acceptable; hence, non-genuine smiles are perhaps uncalled for.

Yet, arguably, findings from the studies referenced above may not be applicable to children, whose experience of winning and losing is likely to differ from that of adults, not to mention that of professional athletes. Moreover, young children may not have fully mastered cultural display rules. In view of these issues, we have decided to focus on smiles expressed by children in this study. We set out to predict the occurrences of genuine smiles and non-genuine smiles displayed by Dutch and Chinese children after they have won and lost in a game.

2. PRODUCTION EXPERIMENT

In a production experiment, we elicited nonverbal expressions associated with winning and losing from Dutch and Chinese children. Our goal was to identify factors that could predict genuine smiles and non-genuine smiles respectively. First, in line with
previous research on the effect of sociality on smiles, we hypothesised that social context would predict occurrence of smiles, such that children who played in pairs would display more genuine smiles as well as non-genuine smiles than children who played alone. Second, we hypothesised that game outcome would also be a significant predictor of smiles. More specifically, we expected more genuine smiles among winning than losing children, and more non-genuine smiles among losing than winning children. In addition, we expected more non-genuine smiles from Chinese children who lost in pairs than Dutch children in the same condition.

2.1. Participants

Thirty-one Dutch (23 male) and 55 Chinese (34 male) children took part in our study, with the informed consent of their parents and teachers. The Dutch participants were 8-year-olds attending a school in Tilburg, the Netherlands. The Chinese participants were 8-year-olds attending a school in Suzhou, China.

2.2. Task

We employed a game paradigm [9] to elicit expressions associated with winning and losing. The game was a simple number guessing game consisting of six rounds. In each round, six cards were displayed on a computer screen but only the number on the first card was visible at first. Children were informed that the numbers ranged from 1 to 10, and were instructed to guess whether the number on the next card would be higher or lower than that currently shown. To win the round, all guesses within the round had to be correct.

To ensure that every child would experience both wins and losses, we manipulated the number on the last card in each round. In half of the rounds, the number on the last card was a likely number given the preceding number (e.g., 5 preceded by 8), as children were told that the numbers ranged from 1 to 10. In the other half, the number on the last card was unlikely given the preceding number (e.g., 10 preceded by 8). This manipulation resulted in at least two wins and two losses for all children.

2.3. Procedure

Children were randomly assigned to play the game alone (15 Dutch, 13 Chinese) or with a classmate of their choice (16 Dutch pairs, 21 Chinese pairs). All children assigned to play in pairs chose a classmate of the same gender; they were free to discuss with their partner as to what guess to make. Children were seated in front of a computer screen and were videotaped throughout the experiment by a camcorder placed on the computer screen. At the start of the game, an experimenter gave instructions in the native language of the children.

For each child or pair, we only retained recordings obtained from the second win and the second loss. This was done to exclude boredom or familiarity with the game as explanations for the nonverbal expressions observed in the recordings, as well as to standardise the recordings across children. This selection resulted in two recordings per child or pair: 56 recordings from children who played alone (30 Dutch, 26 Chinese) and 74 from pairs (32 Dutch pairs, 42 Chinese pairs).

The video recordings showed a frontal view of the children roughly from their shoulders upward. Recordings obtained from pairs were cropped in size such that only one child was visible. The recordings were also trimmed in length such that they depicted the moment at which children had just made their guess, up to the moment at which their response to seeing the game outcome had subsided. The audio track was removed from the recordings, as the game produced sounds that signalled whether the child had won or lost, the information of which might render the subsequent coding biased.

2.4. Results

For each recording, whether or not a smile was present was checked for. In cases where a smile was present, the smile was coded as genuine or non-genuine on the basis of Facial Action Coding System (FACS; [1]). The coding gave rise to two dichotomous variables per recording, one pertaining to the presence of genuine smiles, the other to non-genuine smiles. Overall, a genuine smile was observed in 27.69% of all recordings, whereas a non-genuine smile was observed 33.85% of the time. Logistic regression analyses were conducted to examine if social context (playing alone vs. in pairs), game outcome (winning vs. losing), and culture (being Dutch vs. Chinese) could predict the presence of genuine and non-genuine smiles, respectively.

2.4.1. Genuine smiles

The overall fit of a model with social context, game outcome, and culture as predictors of genuine smiles was significant, $\chi^2 (3) = 36.87, p < .001, R^2 = .36$. In line with our hypotheses, game context and social context turned out to be significant; culture was not predictive of genuine smiles on its own, or in interactions with other predictors. Therefore, we removed culture from the model and re-ran the analysis. The fit of the new model was good, $\chi^2 (2) =$
Contrary to the literature on cultural display rules and our hypothesis, culture did not predict the occurrence of smiles. [5] showed that children understand verbal display rules (e.g., saying thank you) better than they understand facial display rules (e.g., smiling). Thus, cultural differences might exist in the verbal domain but not in smiles. Assessing children’s knowledge and endorsement of verbal display rules would complement our current findings. Moreover, in preparing the recordings for coding, we have noticed that a few children who played in pairs talked to their partner while smiling. Studying the speech that accompanied the smiles would give insights into the interplay between the acoustic and visual information of smiles.

4. REFERENCES


1 Due to the limited length of this extended abstract, we could not report the details of our modelling process here. In sum, the final model reported here performed the best from both a statistical and a substantive perspective.
LAUGHTER DETECTION FOR ON-LINE HUMAN-ROBOT INTERACTION

Marie Tahon¹, Laurence Devillers¹,²

¹LIMSI-CNRS, Human-Machine Communication Department, 91403 Orsay, France
²University Paris-Sorbonne IV, 28 rue Serpente, 75006 Paris, France

marie.tahon@limsi.fr, laurence.devillers@limsi.fr

ABSTRACT

This paper presents a study of laugh classification using a cross-corpus protocol. It aims at the automatic detection of laughs in a real-time human-machine interaction. Positive and negative laughs are tested with different classification tasks and different acoustic feature sets. F-measure results show an improvement on positive laughs classification from 59.5% to 64.5% and negative laughs recognition from 10.3% to 28.5%. In the context of the Chist-Era JOKER project, positive and negative laugh detection drives the policies of the robot Nao. A measure of engagement will be provided using also the number of positive laughs detected during the interaction.

Keywords: Laughter detection, acoustic features, laughter emotional valence

1. INTRODUCTION

The authors focus on laughter detection in human-machine interactions. During real-life affective interactions, many different sounds are collected by the acoustic sensors: neutral and emotional speech, human and non-human sounds. Among human-sounds there are affect bursts: laughs, coughs, cries, etc. In the present study, the authors focus on laughter detection only. In the framework of the Chist-Era JOKER project, positive and negative laugh detection drives the policies of the robot Nao.

Analyzing laughter is a rather complex task because there exists many different types of laughter. In spontaneous speech, most of the laughs overlap speech, so-called “speech-laughs” [12]. Affect bursts detection and particularly laughs detection are of importance for emotion recognition in spontaneous interactions [6] and specifically in interaction using humor strategies like in the JOKER project. Laughs can support positive feelings (joy, amusement, etc.) but also negative affective states (such as contempt [10], sadness or embarrassment). A perceptive test carried on almost 50 isolated laughs have shown a clear difference between laughs perceived as positive or negative in a call center corpus [7]. Several studies [1, 3] found that fundamental frequency, instance duration energy and formants are relevant for clear and well-identified laughs (i.e. “sounds which would be characterized as laughs by an ordinary person if hears in everyday circumstances”). Gaussian Mixture Models have been used for training PLP features [13]. More recently, 13 MFCC trained with HMM have been used for filler/laughter/speech/silence segmentation [9]. At the present time, very few real-life laughs databases are available. In the SEMAINE emotional database [8] and the SSPNet Vocalization Corpus [14], laughter information have been extracted. In a previous study, clear laughs have also been extracted from a spontaneous child-robot interaction [11]. In the JOKER project, an emotional and affect bursts corpora was collected with the aim to train affective models which are described in this paper. The models for detecting laughs are tested in different contexts with data collected in two previous projects ARMEN [4] and JOKER [2]. Our main goal is to present a cross-corpus analysis of laugh.

In this paper, the authors present their study on positive and negative laughs recognition in spontaneous human-machine affective interactions. The section 2 summarizes the databases used for training and testing laughter. Methodology and experimental protocols are presented in section 3. And section 4 presents the conclusions of the study and further works.

2. DATABASES

In this section, three emotional speech and affect bursts databases are presented: one for training and two for testing in cross-corpus conditions.

JOKER training database The training database was collected with two scenarios - jokes and emotion game - which were written in order to elicit emotional speech and laughs. 8 speakers were recorded with a high-quality microphone during an interaction with the robot Nao. In the joke scenario, the robot tells jokes in the aim to elicit laugh. In the emotion game scenario, the speaker has to act emotions (anger, sadness, happiness or neutral state) so as to be recognized by the robot. The recordings collected during the interaction contains emotional speech and affect bursts: laughs but also noise (most of noise sounds being microphone noise or striking table), cough and blows (breathing or blowing). In order to have balanced classes, the robot were asking the speaker to act affect bursts at the end of each scenarios. Each recording has been segmented and transcribed, the number of segments per emotional class and affect bursts is summarized in table 1.
ARMEN test database: The ARMEN corpus [4] was collected in order to collect spontaneous emotional speech with dependent people interacting with a Virtual Agent. In a first phase, the subject was invited by the interviewer to act emotions on purpose, by exaggerating the emotional tone of his voice. In the second phase, the subject would interact with the dialog system designed to induce emotions by projection: a daily situation with an emotional potential. 77 French participants (48 men and 29 women) from 18 to 90 year-old were recorded for a total duration of about 70 min. of speech. Laughs have been perceptively annotated in context, the number of laugh segments is summarized in table 1.

JOKER test database: The JOKER testing database [2] have been collected during funny human-robot interactions. This database was collected to study correlations between Nao’s jokes (social, excessive, self-derision jokes and serious questions) and the user’s appreciation. 18 young adults were recorded. Only laughs have been segmented yet (see table 1).

### Table 1: Number of segments per class for each database, total and mean duration in sec.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Total duration</th>
<th>Mean duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>128</td>
<td>240.7±1.88</td>
</tr>
<tr>
<td>Pos. Laugh.</td>
<td>140</td>
<td>139.2±0.99</td>
</tr>
<tr>
<td>Neg. Laugh.</td>
<td>117</td>
<td>60.9±0.52</td>
</tr>
<tr>
<td>Blow</td>
<td>140</td>
<td>117.7±0.84</td>
</tr>
<tr>
<td>Cough</td>
<td>85</td>
<td>65.0±0.77</td>
</tr>
<tr>
<td>Anger</td>
<td>140</td>
<td>263.7±1.88</td>
</tr>
<tr>
<td>Sadness</td>
<td>140</td>
<td>190.1±1.79</td>
</tr>
<tr>
<td>Neutral</td>
<td>144</td>
<td>249.8±1.73</td>
</tr>
<tr>
<td>Happiness</td>
<td>140</td>
<td>237.5±1.70</td>
</tr>
</tbody>
</table>

| JOKER-pos  | 48             | 93.1±1.94     |
| JOKER-neg  | 27             | 33.2±1.23     |
| ARMEN-pos  | 226            | 252.9±1.12    |
| ARMEN-neg  | 227            | 20.4±0.76     |

Preliminary analysis on acoustic features show that negative laughs are often confused with positive laughs are blows, then it must be relevant to perform hierarchical binary classification tests. Two different configurations are tested (table 2) one is parallel, one is hierarchical.

### Table 2: Parallel (up) and hierarchical (down) configurations.

<table>
<thead>
<tr>
<th>CV</th>
<th>Parallel</th>
<th>Hierarchic</th>
<th>HieraMix</th>
</tr>
</thead>
<tbody>
<tr>
<td>pos</td>
<td>Mean ± SD</td>
<td>Mean ± SD</td>
<td>Mean ± SD</td>
</tr>
<tr>
<td>neg</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallel</td>
<td>67.8 ± 59.3</td>
<td>72.7 ± 60.8</td>
<td>75.1 ± 54.7</td>
</tr>
<tr>
<td>Hierarchic</td>
<td>63.0 ± 56.8</td>
<td>68.8 ± 15.4</td>
<td>61.7 ± 15.2</td>
</tr>
<tr>
<td>Set32</td>
<td>75.1 ± 54.7</td>
<td>61.5 ± 27.4</td>
<td>64.3 ± 28.5</td>
</tr>
<tr>
<td>Set93</td>
<td>66.4 ± 28.5</td>
<td>70.7 ± 15.2</td>
<td></td>
</tr>
<tr>
<td>Set94</td>
<td>67.8 ± 59.3</td>
<td>72.7 ± 60.8</td>
<td></td>
</tr>
<tr>
<td>Set294</td>
<td>65.0 ± 42.5</td>
<td>61.7 ± 27.4</td>
<td></td>
</tr>
<tr>
<td>Set94</td>
<td>67.8 ± 59.3</td>
<td>72.7 ± 60.8</td>
<td></td>
</tr>
<tr>
<td>Set294</td>
<td>65.0 ± 42.5</td>
<td>61.7 ± 27.4</td>
<td></td>
</tr>
<tr>
<td>Set94</td>
<td>67.8 ± 59.3</td>
<td>72.7 ± 60.8</td>
<td></td>
</tr>
<tr>
<td>Set294</td>
<td>65.0 ± 42.5</td>
<td>61.7 ± 27.4</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Results of the classification test (F.measure in %)

3.2. Results

The results are given in terms of F.measure for positive and negative laughs classes only. Cross-validation (CV) and testing (TEST) results are given in table 3. The True (resp. False) Positive, TP (resp. FP) results for the Hierarchic configuration are obtained using the following equations where $P(x|y)^m$ corresponds to the proportion of $y$ laughs classified as $x$ laughs by the model $m$.

\[
TP_x = P(+\mid +)_{\text{laugh}} \times P(+\mid +)_{\text{speech}}
\]

\[
FP_x = P(+\mid +)_{\text{laugh}} \times P(-\mid -)_{\text{speech}}
\]

\[
FP_x = P(+\mid +)_{\text{laugh}} \times P(+\mid -)_{\text{speech}}
\]

\[
FP_x = P(+\mid +)_{\text{laugh}} \times P(+\mid -)_{\text{speech}}
\]

Since the CV results with the parallel configuration are slightly better with Set93 (pos:72.7%; neg: 63.0%) than with Set294 (pos: 67.8%; neg: 59.3%), our feature selection is validated (table 3). One of the advantage of the Hierarchic configuration is to adapt the feature set to the classification task. The last line (Hierarchic) corresponds to a speech classification with Set94 and affect/blow/laugh classifications with Set32. This optimization may also helps to avoid over-fitting since the
speech classes contain around 500 instances and the affect classes contain only 140 instances. The results obtained with all the laughs are best with the Set93, whereas they are better with the HieraMix with positive laughs only. The results obtained with different classification tasks are better since they significantly improve the negative laughs classification.

4. CONCLUSION

In this study, laugh classification is presented. The authors studied different configurations and different acoustic feature sets for classifying laughs. Experiments are carried in a cross-corpus protocol, which means that training and testing acoustic conditions, tasks, speakers are not similar. F-measure obtained results are promising: a combination of acoustic sets improved the recognition of positive laughs from 59.5% to 64.5% and negative laughs from 10.3% to 28.5%. Some technical improvements can be realized: normalization, feature selection, real-time experiments, etc. This experiment also shows that negative laugh are very difficult to detect, because annotations are usually done in context, because they are very often confused with either positive laughs or blows. Multi-modal classification could probably helps to improve recognition rates.

5. REFERENCES


1 http://www.chistera.eu/projects/joker
2 http://www.cs.waikato.ac.nz/ml/weka/
RECOGNIZING EMOTIONS IN DIALOGUES
WITH DISFLUENCIES AND NON-VERBAL VOCALISATIONS

Leimin Tian, Catherine Lai, Johanna D. Moore

School of Informatics, University of Edinburgh
s1219694@sms.ed.ac.uk, clai@inf.ed.ac.uk, j.moore@ed.ac.uk

ABSTRACT

We investigate the usefulness of DISfluencies and Non-verbal Vocalisations (DIS-NV) for recognizing human emotions in dialogues. The proposed features measure filled pauses, fillers, stutters, laughter, and breath in utterances. The predictiveness of DIS-NV features is compared with lexical features and state-of-the-art low-level acoustic features.

Our experimental results show that using DIS-NV features alone is not as predictive as using lexical or acoustic features. However, adding them to lexical or acoustic feature set yields improvement compared to using lexical or acoustic features alone. This indicates that disfluencies and non-verbal vocalisations provide useful information overlooked by the other two types of features for emotion recognition.

Keywords: emotion recognition, dialogue, disfluency, speech processing, HCI

1. INTRODUCTION

Emotions are vital in human cognitive processes. Emotion recognition has long been a focus in human-computer interaction research. State-of-the-art approaches for improving performance of emotion recognition often focus on identifying better feature representations. In this work, our goal is to identify knowledge-driven features that can improve recognition performance.

Psycholinguistic studies have shown that emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention [9]. This in turn influences speech processing and production, which may result in disfluencies and non-verbal vocalisations. Therefore, we would like to investigate the usefulness of DISfluencies and Non-verbal Vocalisations (DIS-NV) for recognizing emotions in dialogues.

One of the most predictive feature sets identified for emotion recognition is the set of acoustic features based on low-level descriptors (LLD). However, in our previous work [7] on the AVEC2012 database [8] of spontaneous dialogues, DIS-NV features were more predictive than acoustic or lexical features for recognizing emotions. We would like to study whether our DIS-NV features remain predictive when the data contains both non-scripted and scripted dialogues. Therefore, we compare our DIS-NV features with LLD acoustic features and lexical features on the IEMOCAP database [1]. Our results show that although DIS-NV features are less predictive than acoustic or lexical features when used alone, they improve performance when combined with existing models.

2. METHOD

2.1. The IEMOCAP Database

The IEMOCAP database contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. Each conversation was about 5 minutes long. There are 10037 utterances in total, of which 4782 utterances were not scripted. When collecting the non-scripted dialogues, the actors were instructed to act out emotionally intense scenarios, e.g., telling a best friend that (s)he has been accepted into his/her most desired university.

Emotions were annotated at the utterance-level with a 1 to 5 integer score of the Arousal (activ-ness), Power (domination), and Valence (positive or negative) emotion dimensions. The mean score over all the annotations was used when the annotators disagreed with each other. We categorized the scores into three classes (<3, =3, >3) to have a clearer view of the relation between emotions and features, and to reduce the influence of imbalanced classes.

2.2. Features

2.2.1. The DIS-NV Features

We studied 5 types of disfluencies and non-verbal vocalisations (DIS-NV): filled pauses (non-verbal insertions, e.g., “eh”), fillers (verbal insertions, e.g., “you know”), stutters, laughter, and breath. We choose them because they are the most common in the data, and they are relatively easy to extract from
transcripts. Disfluencies here refer to interruptions in the flow of speech production. Fluency of speech production may not always be the same with listener’s perception of fluency [6]: Minor disfluencies may be ignored by the listener; In some cases, these tokens could also be perceived as part of a “fluent” utterance (e.g., using a filler at the beginning of an utterance while organizing sentences).

Feature values are calculated as the ratio between the sum duration of each type of DIS-NV and the total duration of the utterance, resulting in 5 DIS-NV features for each utterance. Descriptive statistics of filled pause features are shown in Figure 1 as an example. Utterances containing DIS-NVs are not very frequent in the IEMOCAP database (47.28% in the non-scripted utterances, 24.74% in the scripted utterances). To get a clearer view of value distributions, the statistics shown were computed on a subset of the data which contains all the utterances with disfluencies or non-verbal vocalisations (the DIS-NV subset).

Figure 1: Statistics of filled pause features.

2.2.2. The Lexical Features

The lexical features we extracted are 6 Point-wise Mutual Information (PMI) based features. PMI is a widely used measurement for the relation of words and emotions. It is based on the frequency of a word $w$ having class label $c$, as shown:

$$PMI(c, w) = \log_2\left(\frac{P(c|w)}{P(c)}\right)$$

To calculate PMI values, we first binarized all three emotion dimensions ($<3$, $\geq 3$). PMI values of the scripted and non-scripted data are computed separately. The lexical features we proposed are calculated as the total PMI values of all the words in an utterance for each binarized emotion dimension, resulting in 6 lexical features for each utterance.

Example words with top PMI values are shown in Table 1. In the first column, “A-” represents unaroused, “A+” is excited, “P-” is dominated, “P+” is dominating, “V-” and “V+” represent negativity and positiveness of emotion.

<table>
<thead>
<tr>
<th>Non-scripted data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A-</td>
<td>Academy, Banking, Loan, Numb, Sleep</td>
</tr>
<tr>
<td>A+</td>
<td>Anger, Bloody, Flowers, Freak, Ruined</td>
</tr>
<tr>
<td>P-</td>
<td>Afraid, Beer, Error, Insane, Quit</td>
</tr>
<tr>
<td>P+</td>
<td>Bar, Chick, Duty, F*ck, Mad</td>
</tr>
<tr>
<td>V-</td>
<td>Abuse, B*tch, Die, Iraq, Unfair</td>
</tr>
<tr>
<td>V+</td>
<td>Australia, Cash, Dog, Snow, Tour</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scripted data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A-</td>
<td>Lose, Non, Pets, Skip, Topic</td>
</tr>
<tr>
<td>A+</td>
<td>Bully, Cry, Gods, Jesus, Santa</td>
</tr>
<tr>
<td>P-</td>
<td>Bad, Cliff, Sacrifice, Sneak, Surprise</td>
</tr>
<tr>
<td>P+</td>
<td>Cry, Damn, Lose, Mad, Shut</td>
</tr>
<tr>
<td>V-</td>
<td>Ashamed, Crap, Hell, Sucker, Vile</td>
</tr>
<tr>
<td>V+</td>
<td>Delight, Eating, Gold, Loves, Wish</td>
</tr>
</tbody>
</table>

2.2.3. The LLD Acoustic Features

Our LLD acoustic features were the same as those used in the INTERSPEECH 2010 Paralinguistic Challenge extracted with OpenSMILE [3]. It represents a state-of-the-art feature set for emotion recognition. This feature set has been widely used as a reference for comparing emotion recognition feature sets and classification approaches.

There are 1582 LLD acoustic features, including those extracted by applying functionals (e.g., position of max) to low-level descriptors (e.g., MFCCs, F0, PCM loudness) and their corresponding delta coefficients, the number of pitch onsets, and the total duration of the utterance. Values are computed at the frame-level, with a window size of 60ms and a step of 10ms. Compared to DIS-NV and lexical features, LLD acoustic features overlook global characteristics of the utterance.

2.3. Experimental Settings

Our emotion recognition models were built with the LibSVM [2] classifier using WEKA [4]. We used the C-SVC approach with RBF kernel, and 10-fold cross validation. All features were normalized to [-1,1] before classification. Because of the imbalanced classes, we use weighted F-measure as the evaluation metric.

3. RESULTS AND DISCUSSION

The performance of different feature sets is shown in Table 2. “Mean” in the first row is the un-weighted average of the three emotion dimensions. In the first column, “DN” is the DIS-NV model, “PMI” is the lexical model, “LLD” is the LLD acoustic model.

Our results show that adding DIS-NV features to
Table 2: Performance on the full database.

<table>
<thead>
<tr>
<th>Models</th>
<th>Arousal</th>
<th>Power</th>
<th>Valence</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DN</td>
<td>0.363</td>
<td>0.407</td>
<td>0.328</td>
<td>0.366</td>
</tr>
<tr>
<td>PMI</td>
<td>0.483</td>
<td>0.483</td>
<td>0.332</td>
<td>0.433</td>
</tr>
<tr>
<td>PMI+DN</td>
<td>0.489</td>
<td>0.486</td>
<td>0.406</td>
<td>0.460</td>
</tr>
<tr>
<td>LLD</td>
<td>0.652</td>
<td>0.538</td>
<td>0.535</td>
<td>0.575</td>
</tr>
</tbody>
</table>

Arousal Power Valence Mean

Table 3: Performance on the DIS-NV subset.

<table>
<thead>
<tr>
<th>Models</th>
<th>Arousal</th>
<th>Power</th>
<th>Valence</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DN</td>
<td>0.470</td>
<td>0.453</td>
<td>0.329</td>
<td>0.417</td>
</tr>
<tr>
<td>PMI</td>
<td>0.500</td>
<td>0.467</td>
<td>0.316</td>
<td>0.428</td>
</tr>
<tr>
<td>PMI+DN</td>
<td>0.522</td>
<td>0.475</td>
<td>0.325</td>
<td>0.441</td>
</tr>
<tr>
<td>LLD</td>
<td>0.644</td>
<td>0.523</td>
<td>0.532</td>
<td>0.566</td>
</tr>
<tr>
<td>LLD+DN</td>
<td>0.645</td>
<td>0.525</td>
<td>0.533</td>
<td>0.568</td>
</tr>
</tbody>
</table>

Compared to using the full IEMOCAP database, when using this subset instead, performance of lexical features and LLD acoustic features has a small decrease, while performance of DIS-NV features increases greatly on all emotion dimensions. This verified the negative influence of infrequency of disfluencies and non-verbal vocalisations.

Adding DIS-NV features to lexical feature set remains helpful for all emotion dimensions. Adding DIS-NV features to LLD acoustic features only yields a small gain. The reason may be the great difference between the size of these two feature sets.

We further compared performance of individual DIS-NV features and LLD features with the CFS [5] method, which ranks features based on their individual predictiveness and their correlations with other features. DIS-NV features are always ranked among the top features, especially filled pauses, fillers, and laughter. This indicates that with a better fusion strategy, DIS-NV features may improve performance of LLD features greatly, by highlighting emotionally interesting segments.

Note that DIS-NV and lexical features describe data at the frame-level, while LLD features describe data at the utterance-level. In the future, with advanced fusion strategy that can combine feature sets at different levels with flexible weights, we may be able to combine information contained in these feature sets more efficiently and further boost performance of current emotion recognition models.

4. CONCLUSION

We proposed DIS-NV features measuring disfluencies and non-verbal vocalisations for recognizing emotions in dialogues. We compared their performance with lexical features and state-of-the-art LLD acoustic features. Our experiments on the IEMOCAP database show that using DIS-NV features alone is not enough for building a highly predictive emotion recognition model. However, these features contain information neglected by the lexical or LLD acoustic features. Thus, when fused properly, DIS-NV features may improve performance of current emotion recognition models greatly.

5. REFERENCES

WITH OR WITHOUT WORDS: A PRAGMATIC STUDY OF THE INNOVATIVE QUOTATIVE EXPRESSIONS IN ENGLISH

Sandy Ciroux

Université Libre de Bruxelles (ULB)
sciroux@ulb.ac.be

ABSTRACT

This is a study of the pragmatic features of the innovative quotatives in English. The literature has already deeply analysed the use of these innovative quotatives in the field of sociolinguistics on the one hand, and their development (grammaticalization) in the field of historical linguistics on the other. However, they have not been studied in terms of their pragmatic features yet. One may indeed wonder what the difference between traditional quotatives and innovative quotatives could be. In my research, I found that these innovative quotatives have a tendency to be used to introduce multimodal quotations, i.e. quotations made up with words, body movements and sounds. I hope to be able to demonstrate that these quotatives are used especially to introduce gestures and sounds to permit the utterer to express their own feelings and give their opinion about the quoted element that would be someone’s attitude.

Keywords: Innovative quotatives, pragmatics, gestures, non-lexicalized sounds.

1. INTRODUCTION

A quotative is, in very general terms, a word or phrase used to introduce a quotation. There are two types of quotatives: the traditional quotatives (such as say or think that can introduce direct or indirect speech or thought) and the innovative quotatives. An innovative quotative is a quotative such as be like or go that offers a broader range of possibilities as far as the content of the quote is concerned. In this study, I will explore the content of the quotation introduced by an innovative quotative expression in English with data gained from COCA (Corpus of Contemporary American English)\(^{1}\) and reality shows available on YouTube.

2. METHODOLOGY

As far as COCA is concerned, I chose to work with a series of representative constructions: \([\text{be}] \, \text{like,} \, [\text{pp*}] \, [\text{be}] \, \text{like,} \, [\text{np*}] \, [\text{be}] \, \text{like,} \, [\text{at*}] \, [\text{nn*}] \, [\text{be}] \, \text{like} \) plus all of these with kind of/sorta inserted between \([\text{be}] \) and like.\(^2\) For each construction, I will select a sample of 100 occurrences at random and analyse the content of the quote in terms of nature (thought versus speech and hypothetical versus real) and composition (words, gestures, sounds). The reality shows will permit me to make a more precise analysis because of the sounds and images.

3. PRESENTATION OF MY SUBJECT

It is only relatively recently that researchers such as Clark & Gerrig [4] proposed to see communication as a multimodal phenomenon. They in fact brilliantly argue that in a quotation we can find words but also non-lexicalized sounds (e.g. interjections and onomatopoeias) or even non-verbal sounds (e.g. sighs, laughter and cries) whether or not accompanied by body movements and/or facial expressions. Now that the use of sounds and gestures in quotation is taken into account, what still needs to be studied are the pragmatic features of those innovative quotatives that introduce such non-lexical elements. This is why I am interested in the pragmatic features of these innovative quotatives in English with a particular focus on the importance of gestures, whether bodily (body movements and/or facial expressions) or auditory (non-lexicalized and/or non-verbal). In other words, my research question is “What can an innovative quotative (re)present?”

Innovative quotatives have been studied for some time now, mainly in the field of sociolinguistics ([1], [2], [3] and [5]). What I want to highlight is that their particularity is notably that they can introduce a quoted element that contains either words or gestures or sounds or even a mixture of these three categories. To support this, I analyse videos, in which be like and the like are used. I wish to bring to light the main pragmatic features of the innovative quotatives, i.e. (1) the (re)presentation of gestures (bodily or auditory) by the utterer (2) to allow them to share their feelings and opinion about someone’s attitude, whether hypothetical or real.

\(^{1}\) http://corpus.byu.edu/coca/

\(^{2}\) [\text{pp*}], [\text{np*}], [\text{at*}] and [\text{nn*}] are tags for personal pronoun, proper names, determiners and common nouns on COCA.
5. REFERENCES


ABSTRACT

We present an attempt to describe the semantics of “embarrassment” laughter in aphasic and non-language-impaired discourse based on the samples from the Russian CliPS corpus based on its place in discourse.

Keywords: aphasia, corpus, pear stories, discourse

1. INTRODUCTION

Laughter is a non-verbal vocalization that is known to express positive and negative emotions [1]. A number of studies addressed the type of laughter that could be called “embarrassment” laughter and occurs when the speaker treats the situation as embarrassing or delicate [2,3], for example failure to produce an appropriate word form by people with aphasia [4,5].

We present an attempt to describe the semantics of “embarrassment” laughter in aphasic and non-language-impaired discourse based on the samples from the Russian CliPS corpus.

Russian CliPS (Clinical Pear Stories) corpus is a multimedia corpus of narratives produced by speakers with aphasia and right hemisphere damage, as well as healthy speakers of Russian. The corpus contains retellings of the “Pear story” movie [6].

2. INTERPRETATION OF “EMBARRASSMENT” LAUGHTER

We have analysed 57 narratives (27 stories by people with aphasia and 30 stories by healthy speakers). Only 13 narratives had occurrences of “embarrassment” laughter, seven narratives by people with aphasia and six narratives by healthy speakers.

2.1. Failure to produce an appropriate linguistic form

The embarrassment can be related to failure to produce a linguistic form (example 1) or production of an inappropriate form (example 2) due to linguistic incompetence (examples 1 and 2) mostly in aphasia, or circumstances (example 3). All examples are given in translation from Russian.

(1) And he looks, looks, whoever, how<is he called>, whoever, I don’t know *laugh* [AP-s05, speaker with aphasia]

(2) Three boys one eh grown-up, the second is smaller, the third *laugh* eh well ehm well bigger [AP-s08, speaker with aphasia]

(3) So eh that one continues gathering, *sound of a door creaking* eh a small pause *laugh* [HP-v01, healthy speaker]

2.2. Failure to produce an appropriate interpretation

Laughter can occur when the speaker proposes a certain interpretation of what was happening in the movie and is not sure of the quality of this interpretation (example 4).

(4) He has a strange animal with him. I understood it that way that it is a male_goat, because he doesn’t have an udder,and he bleats like a male goat, well obviously a male_goat *laugh* [HP-v18, healthy speaker]

2.3. Dissatisfaction by the narrative

Laughter in the end of an episode or the story can be explained as the embarrassment of the speaker about the quality of their retelling (example 5).

(5) They left. that’s all. I don’t know *laugh* [AP-s05, speaker with aphasia]

Here we presented a way to interpret the semantics of “embarrassment” laughter based on the place of the laughter occurrence in the discourse. A more thorough analysis will be performed on a larger set of narratives.

This project is supported by the Russian Foundation for Basic Research grant Neural Foundations of Discourse Production: Narrative
Impairments in Brain-damaged Patients (#13-06-00614 A).

3. REFERENCES

Perception of affect in speech by older hearing aid users

Juliane Schmidt 1,2, Diana Herzog 3, Odette Scharenborg 1,4, and Esther Janse 1,4

1 Centre for Language Studies, Radboud University Nijmegen, The Netherlands
2 IMPRS for Language Sciences, Nijmegen, The Netherlands
3 Phonak AG, Stäfa, Switzerland
4 Donders Institute for Brain, Cognition and Behavior, Nijmegen, The Netherlands
{J.Schmidt, O.Scharenborg, E.Janse}@let.ru.nl

1. INTRODUCTION

Normal-hearing listeners make use of visual [4], verbal [7] and prosodic information [1] when interpreting a speaker’s emotional state. As everyday communication settings frequently deprive the listener of visual information (e.g., when not being able to face the speaker), listening to ‘how’ a speaker says something, by picking up on acoustic cues such as pitch, intensity and tempo, may be critical for perceiving affect.

This study investigates how hearing loss affects the perception of the emotion dimensions arousal (aroused/calm) and valence (positive/negative attitude) in older adults using hearing aids. Specifically it is investigated whether wearing a hearing aid improves the correlation between affect ratings and affect-related acoustic parameters. Such a result would indicate that the hearing aid makes users become more sensitive to subtle differences in these parameters.

2. METHODOLOGY

Twenty-three hearing aid users (aged 65+) performed two separate rating tasks for the emotion dimensions arousal and valence in two conditions: with and without wearing their hearing aids. Affect ratings by an age-matched group of 24 adults with age-normal hearing served as a baseline. Participants were Swiss German native speakers. Ratings were collected with a simple 5-step pictorial rating tool [3], ranging from -1 (calm/negative) to +1 (aroused/positive). Stimuli were short and semantically neutral utterances (arousal: 24 utterances, valence: 18 utterances) from an authentic and affectively colored German conversational speech corpus [5]. In a previous study [6], the main acoustic cue for arousal was found to be mean intensity and for valence it was mean F0. These acoustic cues were measured using Praat [2].

3. RESULTS

Statistical analyses were carried out for each emotion dimension separately using linear mixed-effects regression analysis with random intercepts for stimulus and participant. Affective ratings (dependent variable) were compared across conditions (within the hearing aid users: aided vs. unaided) or groups (hearing-aid users vs. reference group) and predicted by the main acoustic parameter for each emotion dimension, i.e., mean intensity for arousal and mean F0 for valence, also allowing for interactions between acoustic parameter and group/condition.

For the hearing-impaired group, the with/without hearing aid condition comparison showed that the main acoustic cue for either emotion dimension was predictive of the affective ratings: Higher mean intensity was interpreted as more aroused ($\beta = 0.077$, $p < .001$); and higher mean F0 was associated with more negative valence ($\beta = -0.005$, $p < .001$). For arousal, hearing-impaired users rated utterances as less aroused without their hearing aids compared to when they were wearing their hearing aids ($\beta = -0.072$, $p < .001$). For valence, there was no difference in affective ratings between the aided and unaided conditions. Importantly, the use of the hearing aid did not interact with the use of the acoustic information for either emotion dimension.

In the between-group analysis, acoustic cues for arousal ($\beta = 0.073$, $p < .001$) and valence ($\beta = -0.005$, $p < .001$) were predictive of the affective ratings. For arousal, hearing-impaired individuals wearing their hearing aid were more sensitive to mean intensity differences than the group with age-normal hearing ($\beta = 0.007$, $p < .05$). For valence, however, ratings by hearing-aid users did not differ from those given by the reference group, and listener group did not interact with the use of the acoustic information.

4. CONCLUSION

With respect to acoustic cues to affect, in line with [6], our results emphasize that mean intensity and mean F0 are important cues for arousal and valence, respectively. While the perception of valence seems to be unaffected by hearing loss, the use of hearing aids does improve the perception of arousal in hearing-impaired older adults.
7. REFERENCES


Notes
Notes
Notes
Notes
Index of Authors

Çakmak, Hüseyin ............................................. 11, 15, 19
Bergelson, Mira B........................................... 45
Ciroux, Sandy ................................................. 43
Curran, Will .................................................. 27
Devillers, Laurence ......................................... 35
Dupont, Stéphane ........................................... 15, 19
Dutoit, Thierry ................................................ 11, 15, 19
El Haddad, Kevin ........................................... 11, 15, 19
Goudbeek, Martijn ........................................... 31
Herzog, Diana .................................................. 47
Hough, Julian ................................................... 23
Janse, Esther ..................................................... 47
Khudyakova, Mariya V ..................................... 45

Kousidis, Spyros ............................................. 23
Lai, Catherine ................................................. 39
McKeown, Gary .............................................. 27
Moinet, Alexis ............................................... 19
Moore, Johanna D ........................................... 39
Mui, Phoebe ..................................................... 31
Nijholt, Anton .................................................. 3
Pelachaud, Catherine ......................................... 5
Sauter, Disa ...................................................... 7
Scharenborg, Odette ........................................ 47
Schlangen, David ............................................ 23
Schmidt, Juliane .............................................. 47
Swerts, Marc .................................................... 31
Tahon, Marie .................................................... 35
Tian, Leimin ..................................................... 39
Trouvain, Jürgen .............................................. 9