LAUGHTER DETECTION FOR ON-LINE HUMAN-ROBOT INTERACTION

Marie Tahon, Laurence Devillers

LAUGHTER DETECTION FOR ON-LINE HUMAN-ROBOT INTERACTION

Marie Tahon[1], Laurence Devillers[1,2]

[1]LIMSI-CNRS, Human-Machine Communication Department, 91403 Orsay, France

[2]University 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[1] project leaded by the LIMSI-CNRS, 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).

Annotation	# inst.	total duration	mean duration
training database			
Noise	128	240.7	1.88
Pos. Laugh.	140	139.2	0.99
Neg. Laugh.	117	60.9	0.52
Blow	140	117.7	0.84
Cough	85	65.0	0.77
Anger	140	263.7	1.88
Sadness	144	190.1	1.79
Neutral	144	249.8	1.73
Happiness	140	237.5	1.70
testing laugh databases			
JOKER-pos	48	93.1	1.94
JOKER-neg	27	33.2	1.23
ARMEN-pos	226	252.9	1.12
ARMEN-neg	27	20.4	0.76

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

3. METHODOLOGY

3.1. Automatic classification and tasks

Automatic classification is performed with the Weka platform[2] with SMO function, RBF kernel and non optimized parameters ($c = 2$ and $\gamma = 0.125$). The complete acoustic set *Set294* contains 294 prosodic, spectral, cepstral, formants and voice quality features extracted on the full, voiced and unvoiced segment [5]. Because the number of acoustic features is more important than the number of instances per class (around 140), a feature selection is essential. Analyzing negative and positive laughs from the training database confirms the state-of-the-art acoustic characteristics: F_0, formants, loudness and spectral features. Two subsets of features were perceptively selected from ANOVA analysis.

- *Set93*: extracted on full segment only
- *Set32*: spectral, formants and F_0 based features

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.

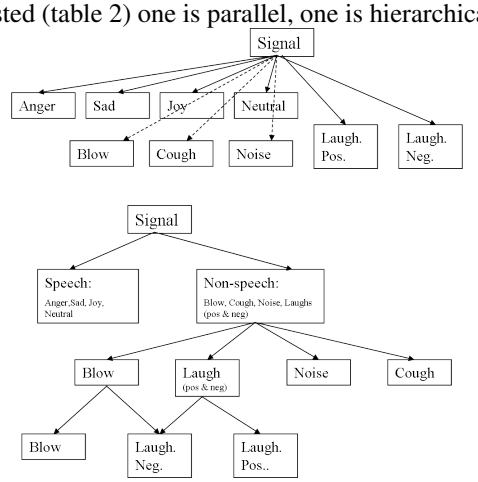

Table 2: Parallel (up) and hierarchical (down) configurations.

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.

$$TP_+ = P(+|+)^{laugh} \times P(+|+)^{affect} \times P(+|+)^{speech}$$
$$TP_- = (P(-|-)^{laugh} + P(-|-)^{blow}) \times P(-|-)^{affect} \times P(-|-)^{speech}$$
$$FP_- = P(-|+)^{laugh} \times P(-|+)^{affect} \times P(-|+)^{speech}$$
$$FP_+ = P(+|-)^{laugh} + P(+|-)^{blow}) \times P(+|-)^{affect} \times P(+|-)^{speech}$$

	CV		TEST	
	pos	neg	pos	neg
Parallel/Set294	67.8	59.3	59.5	10.3
Parallel/Set93	72.7	63.0	60.8	5.6
Hierarchic/Set93			68.8	15.4
Parallel/Set32	75.1	54.7	61.7	15.2
Hierarchic/Set32			61.5	27.4
HieraMix			64.5	28.5

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

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 (HieraMix) 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] Bachorowski, J.-A., Smoski, M. J., Owren, M. J. 2001. The acoustic features of human laughter. *Journal of the Acoustical Society of America* 110 (3), 1581–1597.

[2] Bechade, L., Sabouret, N., Devillers, L. 2014. Automatic construction of behavioral rules from traces of human-robot interaction. *Journées Nationales de la Robotique Interactive*.

[3] Campbell, N. 2004. Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation. *International Conference on Spoken Language Processing* Jeju Island, Korea.

[4] Chastagnol, C., Devillers, L. 2012. Collecting spontaneous emotional data for a social assistive robot. *ES^3 2012 workshop, as part of LREC 2012*.

[5] Devillers, L., Tahon, M., Sehili, M., Delaborde, A. 2015. Détection des états affectifs lors d'interactions parlées: robustesse des indices non verbaux. *Revue Traitement Automatique du Language Parlé, numéro spécial "Oral"* in press.

[6] Devillers, L., Vidrascu, L. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. *Interspeech* Pittsburgh, PA, USA.

[7] Devillers, L., Vidrascu, L. 2007. Positive and negative emotional states behind laughs in spontaneous spoken dialogs. *Interdisciplinary Workshop on The Phonetics of Laughter* Saarbrücken, Germany. 37–40.

[8] McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schröder, M. 2012. The semaine database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, Issue 1, 5–17.

[9] Salamin, H., Polychriniou, A., Vinciarelli, A. 2013. Automatic detection of laughters and fillers in spontaneous mobile phone conversations. *IEEE International Conference on Systems, Man and Cybernetics*.

[10] Schröder, M. 2003. Experimental study of affect bursts. *Speech Communication - Special session on speech and emotion* vol. 40, Issue 1-2, 99–116.

[11] Tahon, M., Delaborde, A., Devillers, L. 2012. Corpus of children voices for mid-level social markers and affect bursts analysis. *LREC* Istanbul, Turkey.

[12] Trouvain, J. 2001. Phonetics aspects of "speech-laughs". *Orality and gestuality (ORAGE 2001)* 634–639.

[13] Truong, K. P., van Leeuwen, D. A. 2007. Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features. *Interdisciplinary Workshop on The Phonetics of Laughter* Saarbrucken, Germany.

[14] Vinciarelli, A., Salamin, H., Polychriniou, A., Mohammadi, G., Origlia, A. 2011. From nonverbal cues to perception: personality and social attractiveness. *International conference on Cognitive Behavioural Systems (COST'11)*. Springer-Verlag Berlin, Heidelberg 60–72.

[1] http://www.chistera.eu/projects/joker

[2] http://www.cs.waikato.ac.nz/ml/weka/