Measuring instantaneous laughter intensity from acoustic features

Jérôme Urbain* and Thierry Dutoit

TCTS Lab, Faculty of Engineering, University of Mons, Belgium

Being able to process and express emotional signals when interacting with humans is an important feature for machines acting in roles like companions or tutors. Laughter is a very important signal that regulates human conversations. It is however hard for machines, which usually have no real comprehension of the phenomenon that triggered laughter, to understand the meaning of human laughs and, in consequence, to react accordingly.

In this paper, we explore one dimension to characterize laughs: their intensity[1]. Without better understanding the conversation, a machine that can infer the intensity of users' laughs will be better equipped to select an appropriate answer (which can be laughing at an intensity related to the detected laugh).

In [1], an online evaluation study has been conducted where naive raters where asked to estimate the intensity of audiovisual laughter clips. They had to provide one intensity value for each laugh, on a 5-point Likert scale ranging from very low intensity to very high intensity. Each episode of the AVlaughterCycle database [2], that contains one thousand laughs recorded with a webcam and a head-mounted microphone, has been rated by at least 6 different participants. Audiovisual features that correlate with the global intensity have been identified. For example, the range of MFCC0 (related to the acoustic energy) and the maximum opening of the mouth over a laughter episode are correlated with the perceived intensity of this episode.

Here, we extend this work by investigating the possibility to automatically draw instantaneous intensity curves. The advantages compared to the global intensity value are the following: 1) the intensity can be obtained in real-time, without waiting for the end of the episode 2) such curves can be useful to drive laughter synthesis 3) instantaneous intensity curves can provide more insight to understand what creates the local and global perceptions of intensity.

To do so, 49 laughs uttered by 3 speakers of the AVLaughterCycle database and ranging across the global intensity values scored in [1] have been manually annotated by one rater. One intensity value was assigned every 10ms, using only the audio signal.

A linear combination of acoustic features has been designed to match the manual annotation. Figure 1 displays the manual intensity for one laugh, together with the intensity predicted from 2 audio features: loudness (i.e. the perceived acoustic amplitude) and F0. The automatic intensity curve is a weighted sum of these two features, followed by median filtering to smooth the output.

*jerome.urbain@umons.ac.be

[1]In this paper, the term *intensity* is used to refer to how amused the laugher seem to be.

We can see that the automatic curve is matching the trend of the manual annotation. Furthermore, the overall laughter intensity can be extracted from the continuous annotation curve: correlation coefficients between the median intensity scored by users and the intensity predicted from acoustic features are over 0.7 for 21 out of 23 subjects[2].

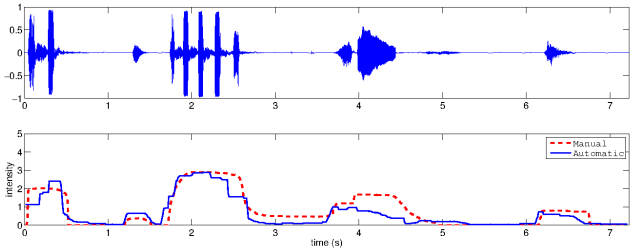

Figure 1: Example of laughter continuous intensity curve. Top: waveform; Bottom: manual and automatic intensity curves.)

Work is in progress to optimize the computation of the continuous intensity (with a trained algorithm using several features instead of the manually designed linear combination) and ensure that the intensity values fall in the same ranges for all the subjects. This is indeed one issue of the weighted sum: it is able to detect which laugh or laughter segment is more intense than another one within one subject (which explains the high correlations given above), but the range of values differ from one participant to the other. To overcome these problems, we are currently training neural networks. The first results are promising both for the continuous and global intensities, but we still have to smooth the continuous curves and perform cross-validation. In addition, we could consider adding visual features to have more robust estimations (in particular for laughs with low acoustic contributions).

Acknowledgment

This work was supported by the European FP7-ICT-FET project ILHAIRE (grant $n°270780$).

References

[1] R. Niewiadomski, J. Urbain, C. Pelachaud, and T. Dutoit. Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases. In *Proc of the ES^3 Workshop, Satellite of LREC 2012*, Istanbul, Turkey, May 2012.

[2] Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. The AVLaughterCycle database. In *Proc. of LREC'10*, Valletta, Malta, May 2010.

[2]The 24$^{\text{th}}$ subject of the AVLaughterCycle database only laughed 4 times, each time with an intensity 1, which prevents us from computing correlations