# On Laughter Intensity Level: Analysis and Estimation

**Kevin El Haddad, Huseyin Cakmak, Thierry Dutoit**
numediart Institute, University of Mons /31 Boulevard Dolez, Mons, Belgium
{kevin.elhaddad, huseyin.cakmak, thierry.dutoit}@umons.ac.be

## Abstract

This work focuses on laughter intensity level, the way it is perceived and suggests ways to estimate it automatically. In the first part of this paper, we present a laughter intensity database which is collected through online perception tests. Participants are asked to rate the overall intensity of laughs. Presented laughs are either audio only or visual only or audiovisual. Statistical analysis show that the perceived intensity is significantly higher when the modality is visual only and suggests that the audio cue might have the biggest influence on laughter intensity perception.We also show that the order by which the modalities are presented to the raters may influence the perception of laughter intensity. In the second part, different estimation/classification techniques were tested including GMM-based mapping and common classification techniques. A set of features were defined, extracted and tested for classification. Results show that the estimation of the global audio laughter intensity is possible with good classification performances.

## 1 Introduction

Laughter is everywhere. So much that we often do not even notice it. It is in common believes that laughter has a strong connection with humor (G. and W., 2015). Most of us seek out laughter and people who make us laugh, and use it in our gatherings and social interactions. Laughter also plays an important role in making sure we interact with each other smoothly. It provides social bonding signals that allow our conversations to flow seamlessly between topics; to help us repair conversations that are breaking down; and to end our conversations on a positive note. In the last decades, with the development of human-machine interactions and various progress in speech processing, laughter became a signal which machines should be able to detect, analyze and produce. This work focuses on the estimation of laughter intensity from acoustic features.

In 2001, Ruch and Ekman (Ruch and Ekman, 2001) published an extensive report on the production of laughter. They investigated various aspects like phonation, respiration, muscular and facial activities. Laughter is described as an inarticulate utterance. Its cycle is around 200 ms and it is usually operated on the expiratory reserve volume. The same year, Bachorowski et al. (Bachorowski et al., 2001) focused on the acoustic properties of human laughter and its differences with speech. They found that laughter yields higher fundamental frequencies than speech, formant frequencies in laughter correspond to central vowels and unvoiced laughter accounts for 40 to 50% of laughter occurrences. Chafe (Chafe, 2007) also describes the mechanical production of laughter and presents various acoustic laughter patterns. A common conclusion of these studies is the high variability of the laughter phenomenon, in terms of voicing, fundamental frequency, intensity and, more generally, types of sounds (grunts, cackles, pants, snort-like sounds, etc.).

Intensity is an important dimension of laughter. The notion of intensity seems so natural that most researchers do not define it (e.g., (Glenn, 2003; Chafe, 2007; Edmonson, 1987)). In (Ruch, 1993), Ruch defines the emotion of exhilaration, which is one of the emotions leading to laughter. He discusses different levels of intensity of this emotion and the corresponding behaviors, from

smile at low intensity to laughter accompanied by posture changes (throwing back the head, vibrations of the trunk and shoulders) at high intensity. Furthermore, intensity is encoded differently by individuals, with reference to their own laughing style (Edmonson, 1987).

Since intensity is a fundamental dimension, frequently and naturally used to describe laughs, it appears as an important feature to be able to estimate for further use in laughter synthesis (Urbain et al., 2014) or recognition. Indeed, it can give valuable information about the state of a participant in a human-machine interaction system. It is also a convenient layer in interactive systems to separate the processes of deciding to laugh (with a target intensity), which is independent from the laughter synthesis voice and style, and synthesizing the corresponding laugh, which obviously depends on the modeled individual traits. In this paper we will use the term *intensity* to refer to the intensity level of the laughter perceived by a listener.

This paper revolves around laughter intensity, how it is perceived and proposes a machine learning based method to detect it.

This paper is organized as follows : Section 2 gives details on the intensity data collection process, Section 3 provides an analysis of the collected data, Section 4 presents a method for laughter intensity level estimation and the experiments leading to it. Finally, Section 5 concludes the paper and give future work perspectives.

## 2 Online perception tests

To collect the intensity data, online tests were conducted. Participants were asked to rate the intensity of laughs on a 5-point scale ranging from 0 to 4. Laughs from 3 subjects were evaluated in this test. Two subjects (1 male, 1 female) from the AVLC Database (Urbain et al., 2010) and one subject (male) from the AVLASYN Database (Çakmak et al., 2014). A total of 334 laughs were used. The number of laughs from each of the subjects is given in Table 1.

Due to relevant data availability, the laughs from the AVLC Database were evaluated only on the audio while laughs from the AVLASYN database were evaluated along 3 different modalities; audio only, video only (video without sound) and both together.

Table 1: Number of laughs for each subject in the experiment

| | |
|---|---|
| AVLC DB (Subject 6) | 67 |
| AVLC DB (Subject 14) | 65 |
| AVLASYN DB (D4) | 202 |
| TOTAL | 334 |

In the case of the AVLASYN Database, 7331 ratings were collected from 226 participants (135 males and 91 females from 18 to 77 years old with an average age of 31.46 and a standard deviation of 10.23).

Table 2 gives the average number of time each file has been evaluated in each of the 3 parts.

Table 2: Average number of time each file has been evaluated in each part of the test

| Modality | Average (std) |
|---|---|
| Audio only | 11.78 (3.47) |
| Video only | 12.03 (3.30) |
| Audiovisual | 11.95 (3.48) |

In the case of the AVLC Database, 1505 evaluations were collected from 40 participants (32 males and 8 females from 20 to 61 years old with an average age of 35.38 and a standard deviation of 10.43). The pipeline followed for the test is the same as above but it contains only one part which is the audio only and each participant was asked to evaluate 40 laughs. Each file has been evaluated 11.40 times on average with a standard deviation of 2.96.

## 3 Data Analysis

### 3.1 Analysis of the perceived intensity in each specific modality

Our first experiment focuses on the possible difference between the perceived intensity with respect to the different modalities in the case of the AVLASYN Database. If we calculate the Pearson's correlation coefficients on the mean intensity values obtained for each single laugh in the different modalities that have been tested, we obtain the following matrix :

$$\begin{bmatrix} & Audio & Video & AV \\ Audio & 1.0000 & 0.9002 & 0.9654 \\ Video & 0.9002 & 1.0000 & 0.9185 \\ AV & 0.9654 & 0.9185 & 1.0000 \end{bmatrix}$$

As expected, there is a strong correlation between the cases. However, we see that the correlation is even stronger between Audio only and Audivisual than between Audio only and Video only. The mean and standard errors of intensity scores of each part are given in Table 3.

Table 3: Mean and standard errors of intensity scores for each part

| Part | Mean (std. err.) |
|---|---|
| Audio only | 1.80 (0.0049) |
| Video only | 2.00 (0.0044) |
| Audiovisual | 1.80 (0.0048) |

This suggests that there might be a difference in the perception of laughter intensity when audio is not present.

#### 3.1.1 Analysis of variance on modalities

To verify this hypothesis, we have conducted an ANOVA test with a post-hoc TUKEY Honest Significant Difference analysis with a confidence level of 99% between the results to the different parts (modalities) of the online test. The pairwise p-values are given below with significant differences in bold :

$$\begin{bmatrix} & Audio & Video & AV \\ Audio & - & \mathbf{0.00} & 0.18 \\ Video & \mathbf{0.00} & - & \mathbf{0.00} \\ AV & 0.18 & \mathbf{0.00} & - \end{bmatrix}$$

The p-values comparison shows that there is a significant difference between the visual only modality and the two others. This confirms our thoughts that the visual modality alone is perceived differently than the case with audio. The mean scores suggest that the visual modality alone tends to be perceived with a higher intensity. Of course, these conclusions are valid only for the studied subject and it might be interesting to investigate the possible generalization of these findings to any laughs or to specific categories of laugh.

#### 3.1.2 Analysis of variance on tests order

One other analysis which may also be interesting is the possible influence of the order in which the modalities are presented to a given participant. As explained above, the perception test is such that 3 different types (audio only, visual only and audiovisual) of files were presented in 3 different successive parts of the test and the order was randomly determined when the test begins. To assess whether or not the order in which the different parts are presented has an influence on the perceived intensity, we perform a One-way ANOVA and the TUKEY HSD post-hoc analysis. P-values are given in Table 4. In this table, for the ease of read, the sequence are defined by 3 numbers. Each number referring to a specific modality ; 1 is for the audio only test, 2 is for the visual only test and 3 is for the audiovisual test. A sequence referred as 123 therefore means that the underlying order of the test was audio only then visual only and finally audiovisual. The main conclusion from this table is that there is a statistically significant influence of the order of the tests on the perceived intensity. It is however hard to find clear patterns from specific test sequences. It is reasonable to think that the position of the video only modality in the test order may have an influence. Indeed, in that modality, the intensity is perceived differently as shown in the previous section.

### 3.2 Analysis of the intensity of each studied subject

Figure 1 gives the boxplots of the intensity values for each subject. We can see that the Subject 14 (female) has the median, 25th percentile and 75th percentile clearly lower than the two other male subjects. The two male subject has similar medians (2.0 and 1.9) and 25th percentiles (both 1.0). However, the 75th percentile is higher for Subject 6 (3.06 against 2.64). Maximum and minimum values are similar for all subject with Subject 6 slightly higher though.

## 4 Audio laughter intensity estimation

Among the possible applications of the intensity information presented in this paper, there is the

Table 4: Pairwise comparison p-values for the different orders in which the test were presented. Significant differences with a confidence level of 95% are given in bold.

| Compared Test Order Pairs | p-values |
|---|---|
| 132-123 | 0.37 |
| 213-123 | 0.20 |
| 231-123 | 0.77 |
| 312-123 | 1.00 |
| 321-123 | 0.17 |
| **213-132** | **0.00** |
| **231-132** | **0.01** |
| 312-132 | 0.63 |
| **321-132** | **0.00** |
| 231-213 | 0.91 |
| **312-213** | **0.04** |
| 321-213 | 1.00 |
| 312-231 | 0.38 |
| 321-231 | 0.89 |
| **321-312** | **0.03** |

estimation of the intensity of a given audio laughter file. It is also important to note that, as shown in this paper, there is not a statistically significant difference between the perceived intensity of an audio only laughter and the same audiovisual laughter. Therefore, we can estimate the intensity of a given audiovisual laugh based on the acoustic information only.

To do this, we propose here to use a Gaussian Mixture Model (GMM) based approach. First, silences are removed from the input audio laughter files. Then, a set of features are extracted from these files and the features that are the most correlated with the output intensity levels are kept. The selected features are then used to train GMMs with full covariance matrices. Doing so, we can model the relationship between the input acoustic features and the corresponding intensity levels. The GMM mapping framework used in this work was first introduced in 1996 by Stylianou (Stylianou, 1996) for voice conversion. The implementation used here is the one of Kain (Kain, 2001) also used in recent work such as (Hueber et al., 2011).

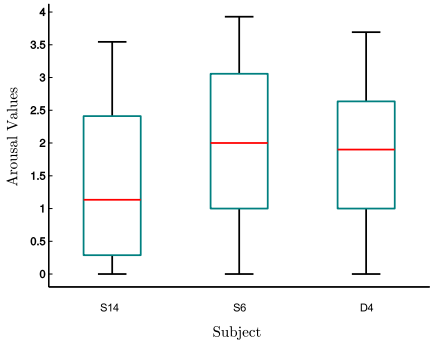


Figure 1: Boxplots for each studied subject. The median is given in red inside the boxes, 25th and 75th percentiles are the limits of the main boxes and the upper and lower tails give the minimum and maximum values of the distribution.

### 4.1 Feature selection

A set of features are extracted from the audio files. Some features are scalar values related to the whole file in the first place while others are continuous features extracted using 10ms windows and 25ms frame shift. The list of extracted features are as follows :

- Spectrogram
- Acoustic Features listed in Table 5 from (Giannakopoulos and Pikrakis, 2014)
- Fundamental Frenquency (F0) extracted using Straight (Kawahara, 2006)

Table 5: List of the 36 features from (Giannakopoulos and Pikrakis, 2014)

| | |
|---|---|
| - Zero Crossing Rate | 1 dim |
| - Energy | 1 dim |
| - Energy Entropy | 1 dim |
| - Spectral Centroid | 2 dim |
| - Spectral Entropy | 1 dim |
| - Spectral flux | 1 dim |
| - Spectral Rolloff | 1 dim |
| - MFCCs | 13 dim |
| - Harmonic Features | 2 dim |
| - Chroma Vector | 12 dim |
| - Spectral Zone | 1 dim |

Since all these features are continuous features, we derived the following descriptors related to the whole file : mean, standard deviation, range, root mean square and histogram values of each feature ; the mentioned histogram values are the number of elements in each of the bins of a histogram calculated on each continuous feature by imposing the number of bins to 3. Among

the most correlated features, we mainly find F0 related features, Chroma vector related features, the mean of the zero-crossing rate and energy entropy standard deviation.

### 4.2 Results

We define 4 different cases of training and testing sets as detailed in Table 6. Case 1 is a training on all the data following a leave-one-out approach. Cases 2 and 3 are used to assess the performances when testing on a subject that was not seen at all in the training. Case 4 is to try if performances are improved when a few examples of the testing subject are shown in training. In this table, the available 3 subjects are named S6 and S14 for subjects 6 and 14 from the AVLC Database and D4 for the subject from AVLASYN Database. Table 7 gives the accuracy results for each case. The accuracy is defined as the number of files for which the intensity estimation error is less than 0.5 (on a scale going from 0 to 4). The table gives all the accuracy values for the cases and sub-cases enumerated in Table 6. We can see that all the accuracy results are over 90% except when the testing is done on the subject S14 (female) for the cases 2 and 3 which correspond to a training on male subjects. We also see that the accuracy increases when we add a few examples of the test subject, even more fore the female subject (see case 4 results).

Table 6: List of train/test sets

| | TRAIN SET | TEST SET |
|---|---|---|
| 1 | - D4+S6+S14 | - leave-one-out |
| 2 | - D4 | - S6<br>- S14 |
| 3 | - D4+S6<br>- D4+S14 | - S14<br>- S6 |
| 4 | - D4+S6+10 files from S14<br>- D4+S14+10 files from S6 | - Remaining of S14<br>- Remaining of S6 |

Table 7: Estimation results for each case

| CASE | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| ACC. (S6) | 96.40% | 94.03% | 91.04% | 92.98% |
| ACC. (S14) | | 81.54% | 86.15% | 90.91% |

Table 8: CASE 1 : Best classification results with $\epsilon = 0.5$ (first row) and $\epsilon = 0$ (second row)

| GMM | TREE | DIS | KNN | NN | SVM |
|---|---|---|---|---|---|
| **96.4%** (46) | 86.3% (12) | 91.6% (44) | 77.2% (51) | 88.6% (81) | **92.9%** (43) |
| **83.8%** (41) | 39.5% (10) | 48.5% (32) | 39.8% (51) | 54.5% (90) | 48.0% (51) |

### 4.3 GMM vs other machine learning methods

In this section we show the reason why the GMM method was chosen by presenting the results of our comparison with other methods. The same training and testing settings as in the previous sections were used here to the binary classification decision tree (TREE), discriminant analysis (DIS), K-Nearest Neighbor (KNN), single layer neural network with 9 neurons (with softmax activation functions and trained with gradient descent) and Support Vector Machine (SVM). To evaluate the classification, we consider that a file is correctly classified with a tolerance ($\epsilon$) of 0.5. This means that if the classification error is at most 0.5 (e.g. 2.5 instead of 2) it is considered as a correct classification. Test were also made with no tolerance in CASE 1 for the sake of comparison. Table 8 shows that GMM has a clear advantage in this respect.

The results for CASE 1 are given in Figure 2 and Table 8. We can see that the best method is clearly the GMM mapping followed by SVM and Discriminant Analysis.

Figure 2: Results for CASE 1 when using the first n most correlated features for training ($n \in [1 : 100]$) and tolerance 0.5

## 5 CONCLUSION AND FUTURE WORKS

In this paper, we studied the intensity level estimation of an audio laughter file from acoustic features. Results show that the estimation is possible. Among the compared methods, GMM-based mapping appears to be the best in all the tested cases. This method also offers good perspectives on the estimation of intensity values not limited to a finite number of classes. Indeed, GMMs can be

used for mapping on decimal values.

In future work we intend to collect to bigger a database of annotated data in order to leverage the power of deep learning. We also intend to link this laughter intensity estimation system with other task such as laughter detection, laughter type classification. This latter will contribute to improve context understanding in intelligent systems.

## 6 Conclusion

This work was focused on laughter intensity. We tried to understand more about its perception by mainly studying the effect of the modality in the perception process. For this we collected a database of laughs, annotated the intensity level of each laugh via online perception tests and analysed them. The results suggested that the audio cue might have the biggest influence on the perception. But further studies are required to confirm it. We compared several machine learning based systems to estimate laughter intensity and showed that the GMMs outperformed the other methods considered. In the future, we intend to increase our database which would allow the use of more advanced techniques such as recurrent and convolutional neural networks.

## References


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

H. Çakmak, J. Urbain, and T. Dutoit. 2014. The AVLASYN database : A synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis. In *Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14)*.

Wallace Chafe. 2007. *The Importance of not being earnest. The feeling behind laughter and humor.*, paperback 2009 edition, volume 3 of *Consciousness & Emotion Book Series*. John Benjamins Publishing Company, Amsterdam, The Nederlands.

Munro S Edmonson. 1987. Notes on laughter. *Anthropological linguistics*, pages 23–34.

McKeown G. and Curran W. 2015. The relationship between laughter intensity and perceived humour. In *The 4th international workshop on laughter and other non-verbal vocalizations in speech*.

Theodoros Giannakopoulos and Aggelos Pikrakis. 2014. *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press.

Phillip J Glenn. 2003. *Laughter in interaction*. Cambridge University Press, Cambridge.

Thomas Hueber, Elie-Laurent Benaroya, Bruce Denby, and Gérard Chollet. 2011. Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. In *INTERSPEECH*, pages 593–596.

Alexander Blouke Kain. 2001. *High resolution voice transformation*. Ph.D. thesis, Oregon Health & Science University.

H. Kawahara. 2006. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6).

W. Ruch. 1993. Exhilaration and humor. *Handbook of emotions*, 1:605–616.

W. Ruch and P. Ekman. 2001. The expressive pattern of laughter. In A. Kaszniak, editor, *Emotion, qualia and consciousness*, pages 426–443. World Scientific Publishers.

Ioannis Stylianou. 1996. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications.

J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. 2010. The avlaughtercycle database. In *Proc. of the Seventh Int. Conf. on Language Resources and Evaluation (LREC'10)*.

Jérôme Urbain, Hüseyin Çakmak, Aurélie Charlier, Maxime Denti, Thierry Dutoit, and Samuel Dupont. 2014. Arousal-driven synthesis of laughter. *Selected Topics in Signal Processing, IEEE Journal of*, 8(2):273–284.