Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases

Radosław Niewiadomski[1], Jérôme Urbain[2], Catherine Pelachaud[1], Thierry Dutoit[2]

[1]LTCI-CNRS Telecom ParisTech
37 rue Dareau, 75014 Paris, France
[2]Université de Mons, Faculté Polytechnique, TCTS Lab
20, Place du Parc - 7000 Mons, Belgium

{niewiado,catherine.pelachaud}@telecom-paristech.fr
{jerome.urbain,thierry.dutoit}@umons.ac.be

Abstract

This paper presents the results of the analysis of laughter expressive behavior. First we present the intensity annotation study of an audiovisual corpus of spontaneous laughter. In the second part of the paper we present analysis of audio and visual cues that influence the perception of laughter intensity, as well as on a study of audio and visual features that differ in laughter inhalation and exhalation phases.

Keywords: laughter, audiovisual synthesis, intensity

1. Introduction

Several research works on social signals were recently undertaken with possible applications in latest HCI technologies such as virtual agents. Laughter is one such signal. It occurs frequently in human-human interaction, and may have many functions and meanings, such as the expression of some emotional states, as well as a social function (Adelsward, 1989). Surprisingly enough, virtual agents - software created to be able to maintain natural multimodal verbal and nonverbal interaction with humans - are still not able to laugh. Knowledge about the expressive patterns of laughter is still limited. Within the long term aim of building a laughing virtual agent, this paper presents the results of our ongoing work on the analysis of laughter expressive behavior. We report on the annotation of an audiovisual corpus of spontaneous laughter, on a study of audio and visual cues that influence the perception of laughter intensity, as well as on a study of audio and visual features that differ in laughter inhalation and exhalation phases.

This paper is structured as follows. In next Section we explain the motivation of this research. Section 3. is dedicated to the description of the intensity annotation protocol. Then, in Section 4. we present the data analysis that we realized so far whereas in Section 5. we present the detailed results. Finally we conclude the paper in Section 6.

2. Motivation for this work

Multimodal laughter synthesis is a complex task. In laughter, the body movements and the tight synchronization between audio and visual signals of the expression is crucial. Laughter is a highly multimodal expression composed of very quick rhythmic shoulders and torso movements, visible inhalation, several facial expressions which are often accompanied with some rhythmic as well as communicative gestures (Ruch and Ekman, 2001). This makes its synthesis particularly challenging. Recent studies on laughter suggest that there exist different types of laughter that can have different expressive patterns (Huber et al., 2009). Consequently, even a small incongruence in laughter synthesis may influence its perception. Particular attention has to be put on the synchronization between modalities which seems to be the key factor of successful laughter synthesis. Thus we need to study first the synchronization between modalities in the human laugh acts.

Even less is known about which audio and visual cues influence the perception of laughter intensity. Differently to many other expressive behaviors studied so far, laughter is a highly multimodal expression. We expect that for laughter the perceived intensity should be a global evaluation that takes into consideration all single monomodal signals. Thus measuring only audio loudness or only mouth openness is not enough to define laughter intensity. Obviously the knowledge about these audio and/or visual cues that influence laughter intensity perception is indispensable in realistic laughter synthesis. In order to properly model laughter in virtual agents, we first need to find the factors that influence the perception of the intensity of human laughs.

In this paper we describe the results of some studies aiming to better understand the expressive patterns of human laughter. We mainly focus on the intensity of laughter. For the purpose of this study we used the AudioVisualLaughterCycle (AVLC) corpus (Urbain et al., 2010) that contains about 1000 spontaneous audio-visual laughter episodes with no overlapping speech. The episodes were recorded with the participation of 24 subjects. Each subject was recorded watching a 10-minutes comedy video. Smart Sensor Integration (Wagner et al., 2009) was used to acquire the signals and manually annotate (and segment) the laughter episodes. The number of laughter episodes for a subject varies from 4 to 82. Each episode was captured with one of two motion capture systems (Optitrack or Zigntrack) and synchronized with the corresponding audiovisual

sample. Each segmented laugh was also phonetically annotated (Urbain and Dutoit, 2011). Two annotation tracks were used: one to indicate the airflow direction (inhaling or exhaling), the other for the actual phonetic transcription.

3. Intensity annotation

We conducted an annotation study of laughter intensity of the AVLC database. The annotation was realized through a web application. This application is composed of a set of web pages; each of them displays one AVLC episode. Our coders were asked to give an overall score of their perceived intensity of the episode using a Likert scale from 1 (low intensity) to 5 (high intensity). Each laugh episode of AVLC was evaluated globally with only one score. There was no obligation to annotate all the available examples (352 episodes). There was no time limitation for the annotation task. Participants could see each sample several times. Once they had evaluated an episode and gone to another one they could not change their previous score. The episodes were displayed in random order. The whole set of episodes was divided into subsets, each of them containing the episodes corresponding to 4 subjects.

For the moment, 2 subsets of the whole database (i.e. 352 out of 995 episodes corresponding to 8 subjects) have been annotated by 15 naive participants mainly from France and Belgium, aged 24-40. Each episode has been annotated by at least 3 and at most 6 coders. Overall agreement between coders was fair: Krippendorff's alpha (Krippendorff, 2012) was .66.

In total we collected 1661 answers. The distribution of the intensity scores in the part of database annotated so far is not uniform. Most of the episodes were evaluated as low intense (see Figures 1 and 2). In more details, the lowest intensity value was used 536 times, score 2 was used 512 times, 3 - 352, 4 - 222, and the maximal score has only been given 39 times.

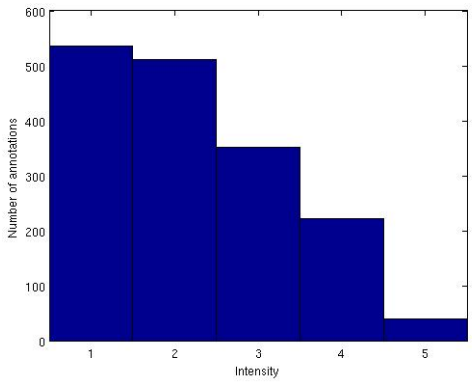

Figure 1: Laughs intensity annotations histogram

4. Data analysis

In this work we focused on two research questions:

- T1) the relation between the perceived intensity and certain audio and/or visual features,

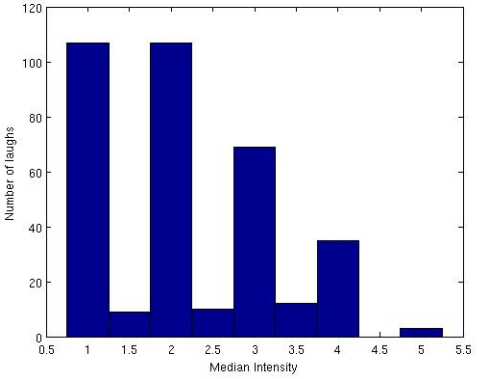

Figure 2: Number of episodes for each degree of intensity

- T2) the relation between the respiration phases and certain audio and/or visual features.

Task T1. The first task relies on the annotation of perceived intensity of laughter (see Section 3.). We aim to discover audio and visual features that correlate with the different degrees of intensity. For each episode we extract several distances between markers that correspond to some action units (Ekman and Friesen, 1978) as well as low-level acoustic descriptors. We are particularly interested in the audio and visual features that can be associated with the intense laughs (such as maximum mouth opening).

Task T2. The second task relies on the annotation of respiration phases in the laughter episodes. Respiration has an important role in the multimodal laughter expression. We expect that information about respiration is crucial to achieve believable audiovisual laughter synthesis: indeed, humans can naturally distinguish these respiration phases when listening or watching to a laugh. The audiovisual signals of the two respiration phases must thus present different patterns. If so, this information can be later used to drive the audio and visual synthesis modules with a common respiration input signal, ensuring the synchronization between the characteristic audio and visual patterns of the two respiration phases. To verify this hypothesis, we analyze the relation between the respiration phases and our audio and visual features and we check if these features take different values in the two respiration phases.

The extracted characteristics are 12 distances corresponding to some facial actions and 58 acoustic low-level descriptors:

- Facial actions are characterized by distances between the markers in the motion capture data. The computed distances correspond to jaw movement (D1), lip height (D2), lip width (D3), cheek raising (D4-5), upper lip protrusion (D6), lower lip protrusion (D7), lip corner movement (D8-9), frown (D10-12). The measurements D4-D5 and D8-D9 roughly correspond to action units considered to be specific for the facial expression of hilarious laughter, namely cheek raising - AU 6 and smile (lip corner up) - AU 12. The remaining measurements correspond to the action units which occurrence in certain laughs is optional or it is still discussed

(Drack et al., 2009) such as AU4 (frowning) or AU 25 (mouth opening) and AU 26 (dropping the jaw). All these characteristics are computed at 25 FPS.

- Acoustic low-level descriptors can be divided into 3 categories: spectral low-level descriptors, measures of the noise level and prosody-related low-level descriptors. Spectral low-level descriptors are 13 MFCCs (as well as their first and second order derivatives), spectral centroid, spectral spread, spectral decrease, spectral flux and spectral variation. Measures of noise are obtained with Harmonic to Noise ratios (HNR, 4 values corresponding to the frequency bands 250-500Hz, 500-1000Hz, 1000-2000Hz and 2000-4000Hz), spectral flatness (4 values also), cepstral peak prominence, chirp group delay and zero crossing rate. Finally, prosody-related low-level descriptors include measures of energy and fundamental frequency. Further details about these low-level descriptors can be found in (Drugman et al., 2011; Peeters, 2003). All these acoustic low-level descriptors were extracted from the 16kHz audio signals, using windows of 512 samples (32ms) shifted by 160 samples (10ms).

For each considered segment (full episode and respiration phase respectively for Task T1 and T2), the frame by frame low-level descriptors (in variable number, depending on the duration of the segment) are mapped to a fixed-length feature vector with the help of the following functionals; minimum over the segment, max, range, mean, standard deviation, skewness, kurtosis, percentage of time spent in the upper quartile ($\%25$), zero-crossing rate (ZCR). Since we had 12 facial distances and 58 acoustic low-level descriptors, we obtain a feature vector of 630 audiovisual features per segment, plus the duration of the segment.

5. Results

We present the results based on the subset of the AVLC corpus for which we have sufficient intensity annotations (see Section 3.). Two subjects had to be removed from the current study due to erroneous motion capture data. Consequently, we had 1336 intensity annotations for the remaining 249 laughs (from 6 subjects).

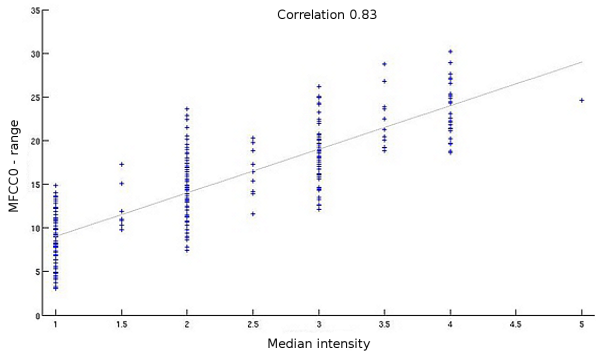

Figure 3: Correlation between median intensity and MFCC0 range

5.1. Intensity and audio visual features

In task 1 we studied the relation between the perceived intensity and several audio and visual features. Concerning the audio features we found strong correlations between several features and the median intensity annotated for each laugh. Spectral features provide the strongest correlations, as well as energy: MFCC0 presents a correlation coefficient (ρ) with the laughter intensity above .8, while loudness is slightly behind. Figures 3 and 4 show the best correlations with the annotated intensity, obtained with MFCC0 range and MFCC2 range, respectively. The detailed data for the 10 best audio descriptors and pitch are presented in Table 1. We can see that the "range" functional is yielding the best correlations for all these lew-level descriptors. Energy descriptors ($MFCC0$, $\Delta MFCC0$, $\Delta\Delta MFCC0$ and Loudness) are the most correlated with laughter intensity, followed by descriptors of the spectral shape (spectral flatness and MFCCs). Pitch, extracted through the ESPS method available in Wavesurfer (Sjölander and Beskow, 2011), is slightly below.

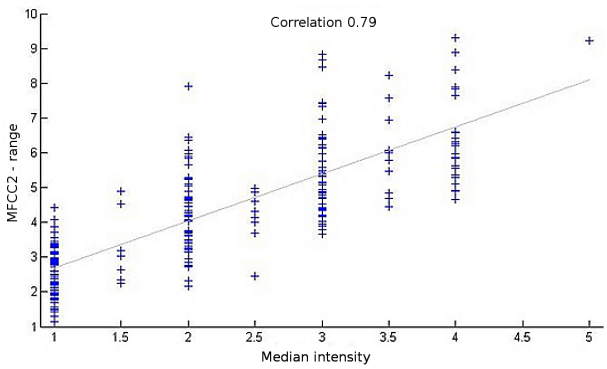

Figure 4: Correlation between median intensity and MFCC2 range

Visual features give slightly lower correlation coefficients. The strongest correlation was observed for the maximum jaw (Figure 5) and lip openings, i.e. the distances D1 and D2, with the "max" functional computed on the whole episode ($\rho = .68$ and .65, respectively).

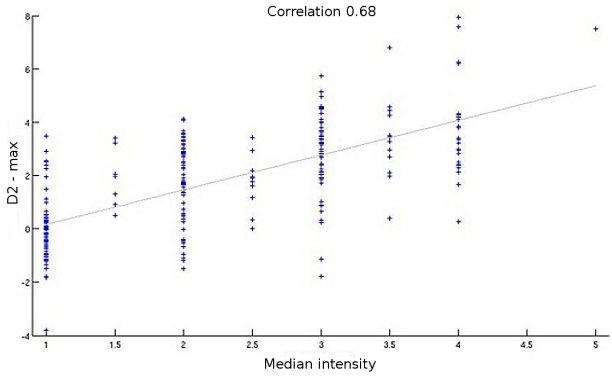

Figure 5: Correlation between median intensity and jaw opening

Table 1: Correlation between laughter median intensity and the 10 best acoustic descriptors (+ pitch)

	SF 1-2kHz	SF 2-4kHz	MFCC0	MFCC2	MFCC5	MFCC8	MFCC10	Δ MFCC0	ΔΔ MFCC0	Loudness	ESPS Pitch
min	-0.77	-0.79	0.20	-0.78	-0.71	-0.59	-0.72	-0.79	-0.75	0.22	-0.02
max	0.23	0.16	0.82	0.36	0.47	0.59	0.54	0.78	0.75	0.78	0.54
range	**0.78**	**0.79**	**0.83**	**0.79**	**0.78**	**0.78**	**0.78**	**0.83**	**0.78**	**0.79**	**0.69**
mean	-0.56	-0.68	0.53	-0.48	-0.32	0.06	-0.11	-0.10	-0.07	0.57	0.30
std	0.66	0.71	0.67	0.69	0.62	0.63	0.68	0.66	0.63	0.69	0.55
skewness	-0.57	-0.57	0.07	-0.45	-0.45	-0.23	-0.39	0.40	-0.22	0.41	0.21
kurtosis	0.44	0.40	0.10	0.25	0.36	0.29	0.31	0.45	0.55	0.41	0.39
ZCR	-0.61	-0.67	-0.22	-0.52	-0.32	-0.43	-0.57	-0.22	-0.27	-0.10	-0.14
%25	0.59	0.62	-0.40	0.20	0.05	-0.02	0.00	-0.49	-0.41	-0.55	0.13

Strong correlation was also observed for maximal lower lip protrusion (D7) ($\rho = .60$). All these three measures received comparable strong correlations when computed as a mean for whole episodes. On the other hand these three distances correspond to the activation of the action units AU 25 and AU 26. This might suggest that the perceived degree of the intensity is correlated with the mean and maximal activation of AU 25/26 and, in other words, with the mouth opening. Similar relations were not observed for other action units that occur in laughter expressions. Indeed, in our test the correlation between the perceived intensity and the measures D4 and D5 was weak ($\rho = .33$ and .43). It suggests that the intensity of the orbicularis oculi activity (i.e. AU6) is not related to the perceived intensity. However it does not mean that this activity was not observed in the dataset. Similarly we did not observe a relation between the measurements corresponding to AU 12 and the perceived intensity. Indeed, the correlation between perceived intensity and the measurements D3, D8, and D9 was only slightly higher (0.33-0.48 for maximum functional, and 0.31 - 0.43 for mean functional) than for the distances corresponding to AU 6. Finally, frowning is even less correlated with the perceived intensity. The observed correlation for the maximal value of the measurement D12 is 0.37. The detailed data are presented in Table 2.

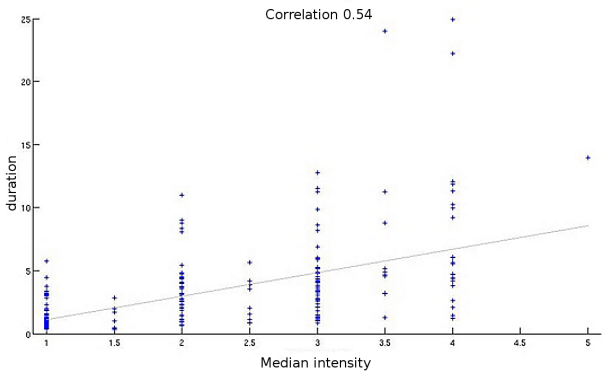

Figure 6: Correlation between median intensity and laughter duration

Interestingly, the overall duration of the laugh is not strongly correlated ($\rho = .54$) with the perceived intensity (Figure 6). In other words, an intense laugh does not necessarily last long, and vice-versa.

These results show us that some audio and/or visual features are strongly related to the perceived intensity of laughs. Hence these features are both good candidates to predict laughter intensity, and helpful to synthesize laughs with the desired intensity.

5.2. Intensity and respiration phases

In task 2 we studied the relation between the perceived intensity and the respiration phases. In total, the 249 laughs contain 419 exhalation phases and 190 inhalation phases. For each feature, we compare its distributions in inhalation and exhalation phases. A Lilliefors test showed that most of the features do not follow a Gaussian distribution; hence a Kolmogorov-Smirnov test was preferred to a t-test to compare the feature distributions over the 2 classes. The Kolmogorov-Smirnov test yielded in highly significant differences in the distributions of the 2 classes, for most of the audiovisual features. Figures 7 and 8 present the distributions, for the two classes, of 4 different features. These experiments illustrate that audiovisual features present different patterns in exhalation and inhalation laughter phases, which confirms our expectations since it is easy for humans to distinguish these phases. These features can be used for segmenting respiration phases in laughter and analyzing their differences.

6. Future works

In this paper we analyzed audio and visual features of spontaneous laughter expressive behavior. First of all we described the intensity annotation of an AVLC audiovisual corpus of spontaneous laughter. We also studied the relation between audio and visual cues of laughter and the perceived laughter intensity, as well as between the audio and visual features and laughter inhalation and exhalation phases.

Several limitations of this work should be noted. First of all the manual annotation of phase respirations can be only roughly done from the audio and/or visual channel. In future we plan to extend our work by using respiration sensor data to increase the segmentation accuracy. Secondly the referred results depend strongly on the choice of

Table 2: Correlation between laughter median intensity and the distances

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
min	0.39	0.32	0.12	0.19	0.11	0.17	0.26	0.15	0.19	-0.08	-0.10	0.08
max	**0.68**	**0.65**	0.48	0.28	0.30	0.20	**0.60**	0.43	0.33	0.12	0.01	0.37
range	0.52	0.48	0.46	0.43	0.44	0.40	0.54	0.3	0.36	0.19	0.15	0.23
mean	**0.64**	**0.61**	0.43	0.26	0.26	0.19	0.54	0.4	0.31	0.04	-0.03	0.29

the episodes, the segmentation method and the context in which the data were collected. Thus, we plan to use data from different video-corpuses to confirm our results. It is particularly important to study the relation between the perceived intensity, some characteristics such as occurrence of AU6 and the type of laughter (social, hilarious). Last but not least the intensity annotation score corresponds to the whole episode but continuous annotation might be more informative as the intensity may not be constant during the laughter episode.

This is an ongoing work. Future works will consist in the more detailed annotation of the existing corpus, more detailed data analysis and finally building laughter models. First of all we plan to extend the intensity annotation of our video-corpus. We will annotate separately the audio and video channels using the same protocol as the one used in Section 3. We are particularly interested in the relation between the evaluation of the single modalities and the overall perception of the intensity. Taking into consideration that laughter episodes are often silent (at least in some phases), this work will give us more knowledge about the role of single modalities in laughter episodes.

Secondly, we are currently investigating the relation between facial actions and the produced laughter sounds, which will also help the synchronized audiovisual laughter synthesis, by looking at the relationship between the annotated vowel-like phones of the AVLC corpus and the shape of the mouth.

Thirdly, after finishing the annotation we discussed with some annotators about the task they had worked on. From these free discussions we observed that our annotators were often trying to evaluate laughter intensity in a subject-dependent way: they evaluated some laughs as relatively intense, i.e. intense when considering that specific person, even if they were not explicitly requested to do so. Our hypothesis is that, while coders may evaluate inter-subject intensity in the first episodes of laughter for a given subject, they rather evaluate the intra-subject intensity when the number of episodes increases. This hypothesis needs to be verified in future works. We ignore this factor in the analysis presented here.

Finally, the results presented here provide new insight for laughter synthesis. We have a better idea of how audiovisual features are related to laughter intensity and respiration phases. We can also use these results for actual prediction of laughter intensity and segmentation of inhalation and exhalation phases.

7. Acknowledgements

The authors would like to thank all volunteer annotators of the database. This work was supported by the European FP7-ICT-FET project ILHAIRE (grant n270780).

8. References

V. Adelsward. 1989. Laughter and dialogue: The social significance of laughter in institutional discourse. *Nordic Journal of Linguistics*, 102(12):107–136.

P. Drack, T. Huber, and W. Ruch. 2009. The apex of happy laughter: A facs-study with actors. In E. Banninger-Huber and D. Peham, editors, *Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions*, pages 32–37. Word Scientific Publisher.

T. Drugman, J. Urbain, and T. Dutoit. 2011. Assessment of audio features for automatic cough detection. In *19th European Signal Processing Conference (Eusipco11)*, pages 1289–1293, Barcelona, Spain, August 29 - September 2.

P. Ekman and W. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.

T. Huber, P. Drack, and W. Ruch. 2009. Sulky and angry laughter: The search for distinct facial displays. In E. Banninger-Huber and D. Peham, editors, *Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions*, pages 38–44. Word Scientific Publisher.

K. Krippendorff. 2012. Computing Krippendorff's alpha-reliability. http://www.asc.upenn.edu/usr/krippendorff/dogs.html (last accessed February 16 2012).

G. Peeters. 2003. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report.

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

K. Sjölander and J. Beskow. 2011. Wavesurfer: open source tool for sound visualization and manipulation [computer program]. http://sourceforge.net/projects/wavesurfer/.

J. Urbain and T. Dutoit. 2011. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In *Proceedings of the fourth international conference on Affective Computing & Intelligent Interaction*, pages 397–406, Memphis, Tennessee, USA. Springer-Verlag.

J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. 2010. The AVLaughterCycle Database. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*,

pages 2996–3001, Valletta, Malta. European Language Resources Association (ELRA).

J. Wagner, E. André, and F. Jung. 2009. Smart sensor integration: A framework for multimodal emotion recognition in realtime. In *Proceedings of the third international conference on Affective Computing & Intelligent Interaction*, pages 1–8, Amsterdam, Holland.

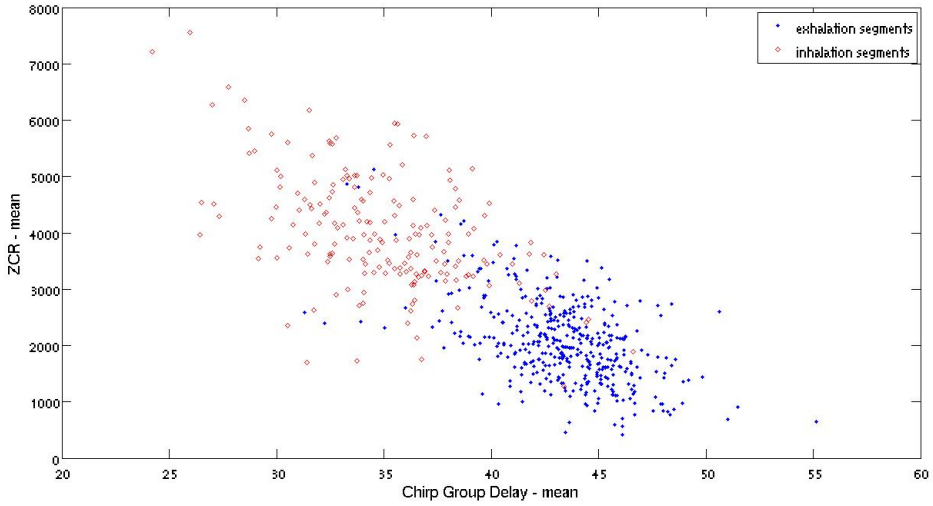

Figure 7: Distribution of mean Chirp Group Delay and mean Zero-Crossing Rate for exhalation and inhaltion laughter phases

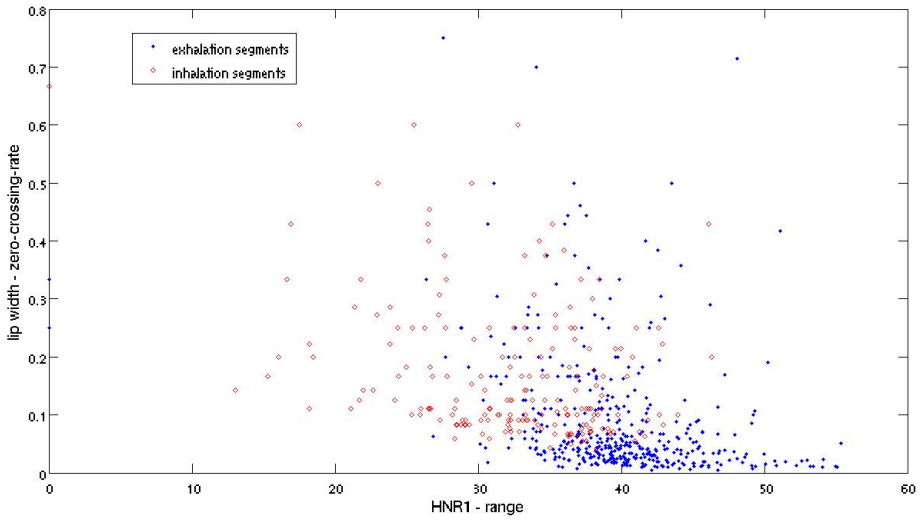

Figure 8: Distribution of mean HNR1 range and zero-crossing rate of AU6 for exhalation and inhalation laughter phases