Analysis of Co-Laughter Gesture Relationship on RGB videos in Dyadic Conversation Context

Hugo Bohy$_1$, Ahmad Hammoudeh$_{123}$, Antoine Maiorca$_1$, Stéphane Dupont$_2$, Thierry Dutoit$_{13}$
[1] ISIA Lab, [2] MAIA Lab,[3] TRAIL
[1,2] University of Mons, Mons, Belgium, [3] Wallonia-Brussels Federation, Belgium
{hugo.bohy, 535653, antoine.maiorca, stephane.dupont, thierry.dutoit}@umons.ac.be

Abstract

The development of virtual agents has enabled human-avatar interactions to become increasingly rich and varied. Moreover, an expressive virtual agent i.e. that mimics the natural expression of emotions, enhances social interaction between a user (human) and an agent (intelligent machine). The set of non-verbal behaviors of a virtual character is, therefore, an important component in the context of human-machine interaction. Laughter is not just an audio signal, but an intrinsic relationship of multimodal non-verbal communication, in addition to audio, it includes facial expressions and body movements. Motion analysis often relies on a relevant motion capture dataset, but the main issue is that the acquisition of such a dataset is expensive and time-consuming. This work studies the relationship between laughter and body movements in dyadic conversations. The body movements were extracted from videos using deep learning based pose estimator model. We found that, in the explored *NDC-ME* dataset, a single statistical feature (i.e, the maximum value, or the maximum of Fourier transform) of a joint movement weakly correlates with laughter intensity by 30%. However, we did not find a direct correlation between audio features and body movements. We discuss about the challenges to use such dataset for the audio-driven co-laughter motion synthesis task.

Keywords: Co-Laughter Motion Analysis, Natural Dyadic Conversation

1. Introduction

The interactive gesture generation task aims to control the gesture of a virtual character with a user control signal. Many works addressed the problem of synthesizing the gesture of an avatar along with a speech modality (Alexanderson et al., 2020; Ahuja et al., 2020). These methods enabled capturing and synthesis of natural co-speech gestures of a virtual character. (Kucherenko et al., 2020) used speech and text jointly as inputs to their proposed model to generate the gestures and reported that the multimodal aspect of their method helps to understand the sentence semantics and outputs natural and diverse gestures. (Yoon et al., 2020) encoded these modalities along with the speaker identity since each expressive behavior highly relies on the speaker.

Nevertheless, motion synthesis from a non-verbal audio input such as laughter is a complex task where no a priori semantic information is available with the audio signal to help with understanding the overall context. However, laughter constitutes an important part of social interaction (McKeown and Curran, 2015) where the smiling and laughing expression of an interlocutor induces a mimicry effect on each partner (El Haddad et al., 2019). The growing interest in virtual environments has led to the development of virtual social agents. The immersive factor of a virtual world is partly induced by the naturalness of the motion of virtual characters. The human-avatar social interaction is an active research topic among the computer vision community and rendering natural motion is a crucial task to enhance the social aspect of the avatar (Garau, 2003). Co-laughter gesture synthesis is thus a relevant task in human computer interaction where it can be exploited in various use cases such as video game development (Mancini et al., 2013) or in a medical context e.g. to enhance the social skills of children with autism spectrum disorder (Didehbani et al., 2016).

The work presented in this paper falls in a wider project aiming at generating co-laughter motion corresponding to the audio given at its input using generative deep neural networks. We present here first analyses results on the relationship between body movements (excluding facial expressions) and several aspects of laughter. These analyses would help us gain a better understanding of our data and thus organize their use to build the previously mentioned generative system. The motion data is not extracted from motion capture sensors but is estimated from the recorded RGB videos directly. Neural networks are powerful tools for learning complex relationships between given modalities within a database. Thus, the proposed analysis allows us to identify whether correlations between laughter, its intensity and the associated movement are significant within a given dataset. If this dataset does not exhibit a high correlation between laughter and body motion, it may be a challenging dataset to train neural networks that synthesize body motion from audio laughter.

This paper is organized as follow: Section 2 reviews the state-of-the-art analysis of the relationship between multiple laughter modalities and co-laughter motion synthesis methods. Section 3 explains the experimental protocol and Section 4 analyzes the experimental results. Section 5 discusses the limitations of this work and proposes some improvements.

2. Related Work

To focus on the synthesis task, it is useful to understand and measure the relationship between laughter as an audio signal and the gesture performed during that laughter. (Griffin et al., 2013) found a significant contrast in the captured motions between different types of laughter (hilarious, social, and non-laughter) and claimed that motion features analysis helped with the classification of laughter type. (Niewiadomski et al., 2016) showed that full-body motion features are sufficient to detect laughter occurrences. (Mancini et al., 2013) pointed out the periodic pattern of the shoulder motion while laughing in the dataset *Multimodal Multiperson Corpus of Laughter in Interaction* (Niewiadomski et al., 2013). (Ishi et al., 2019) focused on laughter intensity to reveal that the degree of smiling face and the occurrences of the front, back, up, and down motions are proportional to the laughter intensity.

(DiLorenzo et al., 2008) proposes a physics-based model to synthesize the torso deformation induces by the air flow while laughing. (Niewiadomski et al., 2014) performs a harmonic analysis of the laughter body motions to get relevant rhythmic features for the generation of body movements. (Ding et al., 2017) synthesized upper body gestures from laughter audio signal based on the captured or defined co-laughter motion correlations. Their approach is based on a statistical framework for head and torso motion and a rule-based method for shoulder motion due to the limitation of their dataset. (Ishi et al., 2019) generated co-speech and laughter motion (eyelids, face, hand and upper body) on physical android robots. The works presented above relied on recorded motion capture datasets of people laughing in multiple contexts. (Jokinen et al., 2016) analyzed videos of social interactions and pointed out the synchrony of body movements with laughter. Similarly, this research aims to identify body motion relationships with laughter from RGB videos and audio signals. However, (Jokinen et al., 2016) estimated bounding boxes around the limbs of the participants.

This work proposes an analysis of the relationship between low-level motion features extracted from RGB videos i.e. the Cartesian position of each joint, the laughter intensity and audio features in the context of a dyadic conversation. This relational study aims to identify any significant correlation between the positions of the joints and the laughter audio signal and intensity. Two approaches are tested and are further explained in Section 3.2.1 regarding the laughter audio signal: first, the audio signal is decomposed into a set of low-level and physical features and then the audio signals are embedded into a latent space from the baseline speech oriented model *Wav2vec 2.0* (Baevski et al., 2020). Finally, the relationship between the 2D Cartesian positions of the skeleton and laughter intensity is established and described in Section 3.2.2.

3. Experiments

3.1. Dataset

In our experiments, we used the dataset *Naturalistic Dyadic Conversation on Moral Emotions* (*NDC-ME*) (Heron et al., 2018). It consists of a collection of dyadic conversations focusing on moral emotions through speaker-listener interactions. In contrast to *IFADV* Corpus (van Son et al., 2008) and the *Cardiff Conversation Database* (Aubrey et al., 2013), the whole upper body of the participants is available in the videos and their motion is not constrained by any object. 21 pairs of participants have been recorded while they were interacting together without following a fixed scenario. The audio and videos have been captured separately. The emotions and the intensity of the expressed emotion of each participant during the recording have been labeled using the annotation tool *ELAN* (Max Planck Institute for Psycholinguistics, 2022) and are available here [1]. The annotation rules follow the protocol [2] used by (El Haddad et al., 2019). The laughter clips are also labeled into 3 categories regarding their intensity: low, medium, and high. At that time, only 7 pairs have been annotated. Following these annotations, the audio and videos in which laughter occurs are extracted from the initial dataset. 186 videos are kept including 10 male and 4 female speakers for a total duration of 199.33 seconds. Then, 2D Cartesian positions of the skeleton joints are extracted from the RGB videos using *OpenPose* (Cao et al., 2018). The skeleton consists of 8 joints representing the upper body of the subject. A frame sample with an estimated skeleton as well as the upper body structure is shown in Figure 1.

3.2. Experimental setup

This part describes the experimental protocol to identify the correlation between the laughter modalities in *NDC-ME* dataset.

Joint movement signals are represented as time series s where $s^i_j = p^i_j - \bar{p_j}$ with p^i_j, the Cartesian position of a joint j at frame i and $\bar{p_j}$ the mean position of the joint j. Thus, s_j is the temporal fluctuations of the position of the joint j around its mean position. Then, the horizontal and vertical component of the motion signal of joint j are respectively noted x_j and y_j. In this work, we consider separately horizontal and vertical movements for the sake of simplicity but it would be interesting to consider both directions. The correlations on

[1] https://zenodo.org/record/3820510

[2] This protocol is available here

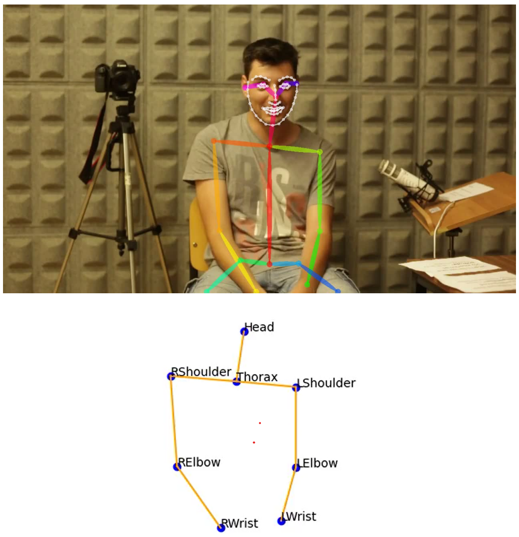

Figure 1: Top: sample of a video with the estimated skeleton and face landmarks. Since this work only focuses on the body skeleton, the face landmarks are ignored. Bottom: structure of the upper body skeleton.

shoulders, elbows and wrists are computed separately for the right and left body parts and we further report the average value.

3.2.1. Body movement and audio features

We wanted to analyse the correlation between the audio signal and the body movement. For the audio signal, we extracted two sets of features per 20 ms frame : one that includes 19 well-known low-level features in the speech analysis domain (3 from LPC, 13 MFCCs and 3 LPCCs), and the other that includes the 512 embedded outputs of the *Wav2vec 2.0* model. For each subset of features, we computed the pearson correlation coefficient between (x_j, y_j) and the time series of audio features.

3.2.2. Body movements and laughter intensity

Firstly, the following features were extracted for each horizontal and vertical joint movement signal (x_j, y_j): In the time domain (power P, maximum amplitude value max, mean value μ and standard deviation σ), and the frequency domain (the maximum value of Fourier Transform $max(FT)$, the mean of Fourier Transform $\mu(FT)$, and peak frequency $f_{pk} = argmax(FT)$). Since laughter videos vary in length, Fourier Transform curves were linearly interpolated in 248 uniform samples between 0 and Nyquist frequency $f_{Nyquist}$. The upper 10% of the frequency range was excluded when finding the peak frequency in order to exclude high-frequency noise ($f_{pk} < 0.9 f_{Nyquist}$). The correlation between those extracted features of joints movement and laughter intensity are then analyzed.

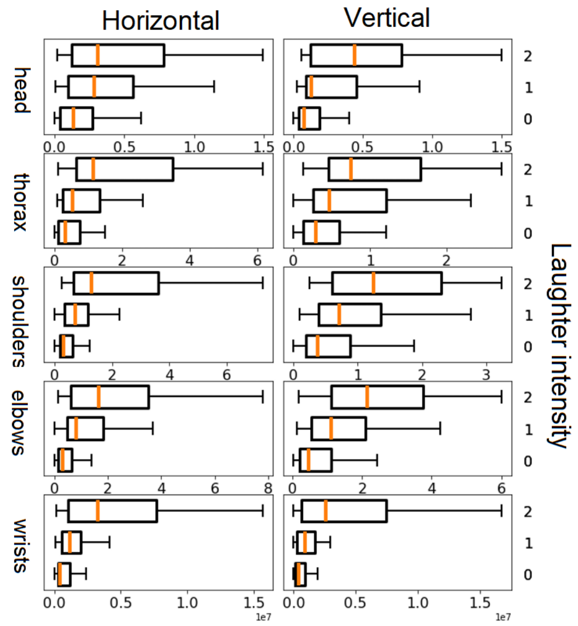

Figure 2: The maximum Fourier transform of a joint movement signal $max(FT(p_j))$ under multiple laughter intensities. Each Row represents a joint and each column represents a direction of movement (horizontal/vertical). Each figure has 3 boxplots (low laughter intensity at 0, medium at 1, and high at 2). The orange line in a boxplot represents the mean.

4. Results

Section 4 presents the results of the correlation analysis between body movements, audio features and laughter intensity.

4.1. Body movements and audio features

Table 1 shows the maximum average correlation between an audio feature and a joint movement. The values depicted informs us about the weak correlation between the evolution of the position of a joint compared to the evolution of an audio feature. However, using embedded features rather than interpretable ones increases the correlation across all joints.

4.2. Body movements and laughter intensity

The correlation between the extracted features and laughter intensity is shown in table 2. Since $max(FT)$ feature has the highest correlation, we visualized the distribution of $max(FT)$ features under multiple laughter intensities in Figure 2. The visualization of $max(FT)$, similar to the other extracted features, resulted in overlapping boxplots. Hence, we conclude that any of the extracted features alone is not sufficient to identify the laughter intensity. However, statistically speaking, the mean value of the distribution (the orange line in Figure 2) increases with laughter intensity.

Feature	**Horizontal Movement**					**Vertical Movement**				
	Head	**Thorax**	**Shoulders**	**Elbows**	**Wrists**	**Head**	**Thorax**	**Shoulders**	**Elbows**	**Wrists**
LPC	0.03	0.02	0.05	0.03	0.02	0.04	0.05	0.07	0.02	0.03
MFCCs	-0.03	-0.01	0.01	-0.01	0.01	-0.08	-0.06	-0.06	-0.04	-0.01
LPCCs	0.05	0.03	0.04	-0.01	-0.01	0.05	0.07	0.07	-0.02	0.08
W2V	**0.09**	**0.08**	**0.07**	**0.08**	**0.09**	**0.11**	**0.09**	**0.09**	**0.10**	**0.09**

Table 1: Maximum average correlation between an audio feature and a joint with respect to its movement direction.

Feature	**Horizontal Movement**					**Vertical Movement**				
	Head	**Thorax**	**Shoulders**	**Elbows**	**Wrists**	**Head**	**Thorax**	**Shoulders**	**Elbows**	**Wrists**
max	0.09	0.25	0.30	0.22	0.26	0.26	**0.39**	**0.25**	0.25	0.20
P	0.08	0.09	0.18	0.10	0.13	0.29	0.25	0.16	0.10	0.10
μ	0.10	0.05	0.06	0.02	0.08	-0.17	-0.19	-0.14	-0.16	-0.15
σ	0.16	0.23	0.28	0.23	0.26	0.35	0.31	0.21	0.27	0.20
$\mu(FT)$	0.13	0.26	0.30	0.24	0.26	0.28	0.37	0.23	0.25	0.18
max(FT)	0.23	**0.32**	**0.36**	**0.32**	**0.34**	**0.36**	0.33	0.24	**0.32**	**0.21**
fpk	**-0.29**	-0.22	-0.20	-0.2	-0.22	-0.22	-0.21	-0.12	-0.20	-0.12

Table 2: Correlation between laughter intensity and a joint movement feature. The power P, maximum amplitude value max, mean value μ and standard deviation σ are computed from the horizontal and vertical motion signals in the time domain. In the frequency domain, the motion features are the maximum value of Fourier Transform $max(FT)$, the mean of Fourier Transform $\mu(FT)$ and and peak frequency f_{pk}. The correlation is bound between -1 and 1. The higher absolute value means a stronger correlation and 0 shows no correlation in the data.

5. Discussion and Challenges

The results presented in Section 4 indicate that, in *NDC-ME* dataset, body movements and audio features seem to be weakly correlated. Further investigation and processing are needed to draw a more robust conclusions. Thus, this dataset seems, at the moment and with this current analysis, challenging for a co-laughter gesture synthesis task. However, we found some aspects in the dataset that might impact the results in our analysis: in some files, speaker speech overlaps with the listener's laughter and we suspect that this influenced the experimental results in Section 4. These need to be removed from the dataset in future work to get more accurate results. One suggestion is the application of channel source separation methods to the audio to distinguish the laughter or speech of each participant and have a better audio representation (more suitable features). Then, the laughter intensity has been subjectively annotated by a single annotator and having a low number of annotators makes the data distribution more sensitive to human error. We suggest to increase the number of annotators and e.g. extracting the mean annotations to reduce this impact. Moreover, since the dataset has not been fully annotated yet, it contains a relatively small amount of laughter examples. Then, in a future work, we would like to extract correlations from audio acoustic features such as pitch or loudness. Moreover, it would be interesting to take into account other modalities such as the type of laughter and the context of the interactions. Finally, in this work, we focus on body movement but face landmarks are available from the *OpenPose* estimation as shown in Figure 1. The relationship between those landmarks and the laughter intensity and laughter audio features can be established in further investigation.

6. Conclusion

This work proposes a method to analyze the relationship between laughter, its intensity and the body movement in recorded dyadic conversations. In contrast with previous works, the gestures are extracted from the RGB videos using a baseline pose estimation method. First, this work highlights around 30% correlation between laughter intensity and motion features where the maximum amplitude of the Fourier transform leads to the highest correlation value. Moreover, the analysis of correlation between interpretable and high-level audio features does not output significant correlation values. This work highlights some of the limitations of *NDC-ME* dataset that we need to take into account in the context of deep generative model training for body motion generation from a laughter audio signal. This analysis opens the way to create datasets suited to build multimodal models that generate the motion of virtual agents from the audio cue.

7. Acknowledgements

This work was supported by Service Public de Wallonie Recherche under grant n° 2010235 - *ARIAC* by *DIGITALWALLONIA4.AI*

8. Bibliographical References

Ahuja, C., Lee, D. W., Ishii, R., and Morency, L.-P. (2020). No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1884–1895, Online, November. Association for Computational Linguistics.

Alexanderson, S., Henter, G. E., Kucherenko, T., and Beskow, J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 39(2):487–496.

Aubrey, A. J., Marshall, D., Rosin, P. L., Vendeventer, J., Cunningham, D. W., and Wallraven, C. (2013). Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2018). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008.

Didehbani, N., Allen, T., Kandalaft, M., Krawczyk, D., and Chapman, S. (2016). Virtual reality social cognition training for children with high functioning autism. *Computers in human behavior*, 62:703–711.

DiLorenzo, P. C., Zordan, V. B., and Sanders, B. L. (2008). Laughing out loud: Control for modeling anatomically inspired laughter using audio. In *ACM SIGGRAPH Asia 2008 papers*, pages 1–8.

Ding, Y., Huang, J., and Pelachaud, C. (2017). Audio-driven laughter behavior controller. *IEEE Transactions on Affective Computing*, 8(4):546–558.

El Haddad, K., Chakravarthula, S. N., and Kennedy, J. (2019). Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles. In *2019 International Conference on Multimodal Interaction*, pages 259–263.

Garau, M. (2003). *The impact of avatar fidelity on social interaction in virtual environments*. University of London, University College London (United Kingdom).

Griffin, H. J., Aung, M. S., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., and Bianchi-Berthouze, N. (2013). Laughter type recognition from whole body motion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 349–355.

Heron, L., Kim, J., Lee, M., El Haddad, K., Dupont, S., Dutoit, T., and Truong, K. (2018). A dyadic conversation dataset on moral emotions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 687–691. IEEE.

Ishi, C. T., Minato, T., and Ishiguro, H. (2019). Analysis and generation of laughter motions, and evaluation in an android robot. *APSIPA Transactions on Signal and Information Processing*, 8.

Jokinen, K., Trong, T. N., and Wilcock, G. (2016). Body movements and laughter recognition: experiments in first encounter dialogues. In *Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, pages 20–24.

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexanderson, S., Leite, I., and Kjellström, H. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

Mancini, M., Ach, L., Bantegnie, E., Baur, T., Berthouze, N., Datta, D., Ding, Y., Dupont, S., Griffin, H. J., Lingenfelser, F., et al. (2013). Laugh when you're winning. In *International Summer Workshop on Multimodal Interfaces*, pages 50–79. Springer.

Max Planck Institute for Psycholinguistics, T. L. A. (2022). Elan (version 6.3) [computer software]. Retrieved from `https://archive.mpi.nl/tla/elan`.

McKeown, G. and Curran, W. (2015). The relationship between laughter intensity and perceived humour. In *The 4th Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech, Enschede, Netherlands*, pages 27–29.

Niewiadomski, R., Mancini, M., Baur, T., Varni, G., Griffin, H., and Aung, M. S. H. (2013). Mmli: Multimodal multiperson corpus of laughter in interaction. In Albert Ali Salah, et al., editors, *Human Behavior Understanding*, pages 184–195, Cham. Springer International Publishing.

Niewiadomski, R., Mancini, M., Ding, Y., Pelachaud, C., and Volpe, G. (2014). Rhythmic body movements of laughter. In *Proceedings of the 16th international conference on multimodal interaction*, pages 299–306.

Niewiadomski, R., Mancini, M., Varni, G., Volpe, G., and Camurri, A. (2016). Automated laughter detection from full-body movements. *IEEE Transactions on Human-Machine Systems*, 46(1):113–123.

van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2008). The ifadv corpus: A free dialog video corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., and Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6).