Building a Multimodal Laughter Database for Emotion Recognition

Merlin Teodosia Suarez, Jocelynn Cu, Madelene Sta. Maria
Center for Empathic Human-Computer Interactions, De La Salle University
2401 Taft Ave., 1004 Manila, Philippines
E-mail: merlin.suarez@delasalle.ph, jiji.cu@delasalle.ph, madelene.stamaria@dlsu.edu.ph

Abstract

Laughter is a significant paralinguistic cue that is largely ignored in multimodal affect analysis. In this work, we investigate how a multimodal laughter corpus can be constructed and annotated both with discrete and dimensional labels of emotions for acted and spontaneous laughter. Professional actors enacted emotions to produce acted clips, while spontaneous laughter was collected from volunteers. Experts annotated acted laughter clips, while volunteers who possess an acceptable empathic quotient score annotated spontaneous laughter clips. The data was pre-processed to remove noise from the environment, and then manually segmented starting from the onset of the expression until its offset. Our findings indicate that laughter carries distinct emotions, and that emotion in laughter is best recognized using audio information rather than facial information. This may be explained by emotion regulation, i.e. laughter is used to suppress or regulate certain emotions. Furthermore, contextual information plays a crucial role in understanding the kind of laughter and emotion in the enactment.

Keywords: laughter corpus, multimodal, emotion analysis

1. Introduction

Laughter is ubiquitous in everyday human encounter, and in the field of affective computing and intelligent behaviour analysis, becomes a significant social cue that contributes to an interaction. Numerous works on laughter analysis involve automatic detection and recognition using audio and video signals (Truong & Van Leeuwen, 2005; Truong and Van Leeuwen, 2007; Petridis & Pantic, 2008; Reuderink, et. al., 2008; Knox, et. al., 2008; Escalera, et. al., 2009; Petridis & Pantic, 2010). Typical computational models of laughter were built using initially audio, and eventually audio-visual information, based on multimodal corpora. Data was collected from various sources: acted data from professional actors and induced using funny videos, comic strips and jokes (Urbain, et. al., 2010), spontaneous data from call centers (Devillers & Vidrascu, 2009) and meeting data (Nachamai & Santhanam, 2008). Usual issues ranged from distinguishing laughter and speech, analysing voiced and unvoiced laughter, and identifying the onset and offset of laughter episodes.

2. Laughter as Social Signal

In this work we built a laughter corpus specifically for the purpose of analyzing the emotion that accompanies a laughter occurrence. We imagine that interactions between users and computing system interfaces can be improved when interfaces are better able to understand the meaning of its user's behavior and emotion. While emotion analysis has typically been achieved via facial expressions and voice, we believe that laughter as a social signal is loaded with meaning that needs to be deciphered and carefully studied. A bigger question needs to be addressed before emotion detection and recognition via laughter can be performed: do humans use laughter to express various emotions (not just happiness, as is obvious), and can laughter in these cases be distinguished from each other? To answer these questions we built a laughter corpus.

3. PinoyLaughter: Multimodal Laughter Emotion Database

To the best of our knowledge, this is the only laughter corpus built for emotion recognition. The PinoyLaughter Database is a multimodal corpus of laughter instances. It is composed of both acted and induced male and female laughter instances.

Acted laughter was collected from two professional actors (one male and one female), who were asked to laugh to express five (5) different emotion labels (roughly translated to their English equivalents, since the actors were given words from the local language to dispel semantic misinterpretation): happiness, giddiness, excitement, embarrassment and hurtful laughter. To arrive at these local emotion labels associated with laughter, linguists conducted focus group discussions to identify emotions that can be associated with laughter in natural interactions.

Figure 1. Data Collection Set-up. The subject is shown wearing a headset with a microphone, and a video camera that records the laughter episodes.

Figure 1 shows the data collection set-up. Each actor was asked to seat in front of a video camera. They were asked to make as little movement as possible so as not to deviate from camera view. A noise-cancelling microphone was used to record audio data. They were asked not to tilt their head in any direction so as to maintain a good frontal view of their face. Each actor was asked to wear dark clothing, and the recording was taken against a white background to minimize noise and pre-processing later on.

To test the quality of the enactments, psychologists were asked to label each enactment. Three (3) male judges who possess graduate degrees in Applied Social and Cultural Psychology were asked to annotate acted laughter. All the judges watched each clip together and discussed among themselves to arrive at one label, including contextual information. In this case, the judges were asked to identify when they think these enactments will occur. In one (1) clip for example, with hurtful laughter, judges identified this occurring when she subject is expressing power over another.

If the emotions are identifiable, then these clips are true representations of the emotion, and can thus be used for laughter analysis. Because the actors were asked to enact a specific emotion, they were interviewed at the end of each enactment to explain the motivation for their expression. These served as contextual information which annotators later relied on to label the emotions.

Manual video segmentation was performed on the recorded data. It was divided into segments using Sony Vegas 9.0 (http://sonycreativesoftware.com). Frames were extracted with a frame rate of 30 frames progressive. Figure 2 shows sample frames taken from a recording.

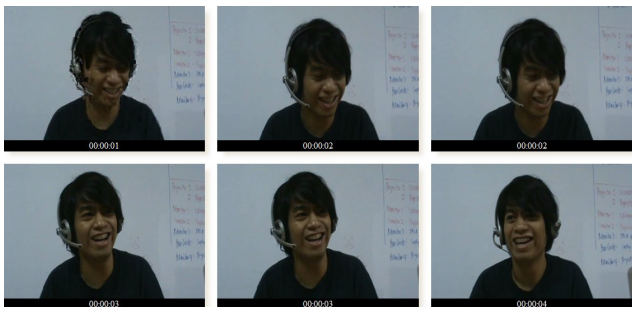

Figure 2. Sample image frames taken from laughter video recording. Frames are extracted at a speed of 30 frames/sec. The changes in head movement are evident in this sequence.

Audio signals were pre-processed were pre-processed by removing signals not related to laughter episodes, such as speed and unnecessary silence, including ambient noise such as static electricity and air-conditioning. Noise-reduced sounds were cut into segments using PRAAT (Boersma and Weenink, 2009). The resulting signals are two-second overlapping segments, with a sample frequency of 44.1 kHz with a window size of 30ms and a sampling frequency of 8 kHz.

A total of 497 audio clips and 3292 images were collected. The active shape model (Chellapa & Zhao, 2006) was used to map the facial points, and the distances between these were extracted. Sixty-four (64) facial points, including corners of the eyes, mouth, lips, nose among others) were taken. Distances between these 64 points were computed, yielding 165 facial distances. In terms of audio features, both prosodic and spectral features were extracted. The prosodic features extracted include the minimum, maximum, mean and standard deviation of the pitch, formants, pitch contour points. The first 13 Mel-frequency cepstral coefficients were extracted as well.

4. Observations and Analyses

Findings indicate that contextual information is relevant in being able to label each clip correctly. Likewise, even while the data was enacted, they were not distinct enough to be labelled easily without contextual information. Laughter is known to be a regulatory mechanism in the expression of emotion (Mesquita & Frijda, 1992). As such, contextual information is necessary for the regulatory processes of suppression and reappraisal to be activated in the expression of emotion. Emotion suppression is typically achieved through an active effort at reducing expressivity (Butler, Lee & Gross, 2007). This may have happened in the observed emotion enactments through laughter. The actors may have used laughter to regulate the felt emotion, i.e. suppress the expression of an emotion. This is why, based on manual observation and experiments, it appears that the voice is the move reliable modality in expressing and recognizing emotions via laughter compared to using the face, as can be seen in Figure 3. Classifiers were constructed, one for the face and another for audio data. The results of these classifiers were weighted (with the assigned weights as shown in the x-axis of Figure 3), and the accuracy scores plotted (as the y-axis). The audio model is a better predictor of laughter compared to the face.

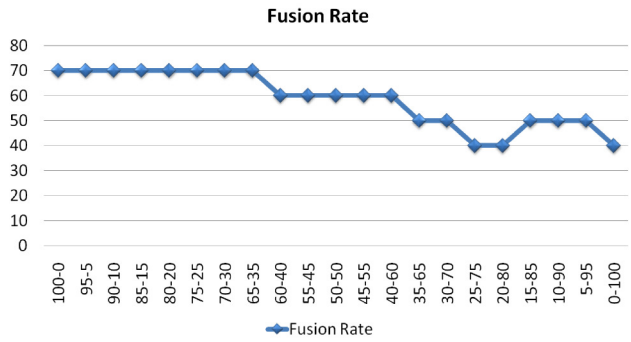

Figure 3. The x-axis shows the combined weights for the voice-face modalities, and the y-axis shows the accuracy score of the machine classifiers. Notice the consistently high accuracy when the voice is given a higher weight than the face modality.

Realizing that acted data cannot build robust models (both for laughter and emotion recognition), induced data was

likewise collected. As in (Nachamai & Santhanam, 2008; Devillers & Vidrascu, 2009; Urbain, et. al., 2010), data was collected by inducing laughter using external stimuli. Data was collected from 3 test subjects. In this case, the subjects were asked to label their emotions using dimensional information of valence and arousal via FeelTrace (Cowie, et. al., 2000), and annotators who passed the empathy quotient test (Baron-Cohen & Wheelwright, 2004) labelled the clips as well. Only those clips with 75% sign agreement were added to the corpus. Another round of annotation was performed on the same clips by the same annotators, this time using discrete labels. The objective of re-annotation was to investigate if it was possible to map discrete labels to a robust cluster in the valence-arousal plane.

Results as shown in Figure 4 indicate that emotions expressed using laughter are consistently found in the positive-valence, high-arousal quadrant. When the local emotion labels were mapped onto Russell's Circumplex Model (Russell, 1980), they were very close to emotions of happiness, delighted and interested.

Interestingly, the emotion label "mapanakit" or hurtful laughter was not elicited properly and therefore not dominant in the corpus. The significance of contextual information in annotating the clips consistently and accurately as found in labelling acted clips was again manifested in labelling these data. The dependence on contextual information is to identify the emotion expressed through laughter suggests an elevated regulatory process in the expression of an emotion. This further suggests that laughter may be an important paralinguistic tool for emotion regulation, particularly in cultures, i.e. collectivist, where there is a stronger need to regulate emotions (Masuda, et al., 2008). Greater suppression is also found to happen among those who live in cultures that highly endorse the expression of positive emotions (Matsumoto et al., 2008; Safdar et al., 2009).

Another interesting results for "mapanakit" are the presence of outliers in the positive arousal, *negative* valence quadrant as shown in Figure 5. We imagine the hurtful laughter clips should score higher in the arousal value, nearer angry rather than afraid. Therefore, additional clips for hurtful laughter need to be collected.

Manual observation of laughter expressing happiness and excitement both exhibit breathing-dominated laughter, moderate pitch and an evident rise and fall of intonation. Laughter expressing giddiness possesses high pitch and energy, with close breathing intervals similar to panting. Hurtful laughter is likewise breathing dominated with a sudden change of emotion at its onset. Embarrassing laughter exhibits long breathes and is high-pitched.

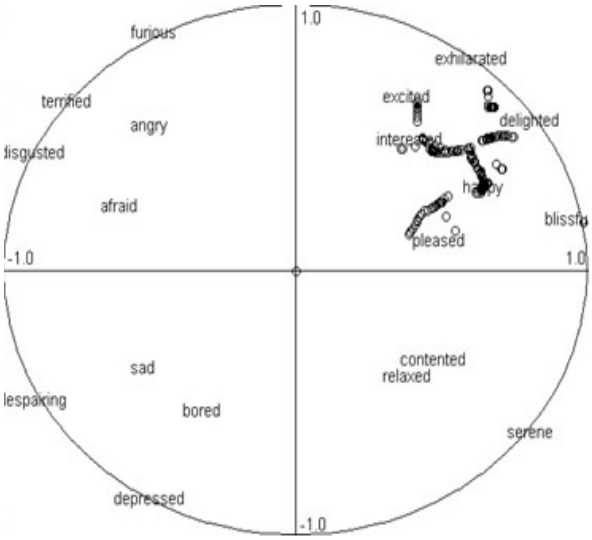

Figure 4. When mapped against Russell's Circumplex Model, laughter instances with label giddiness were consistently found on the upper right quadrant indicating high arousal and positive valence

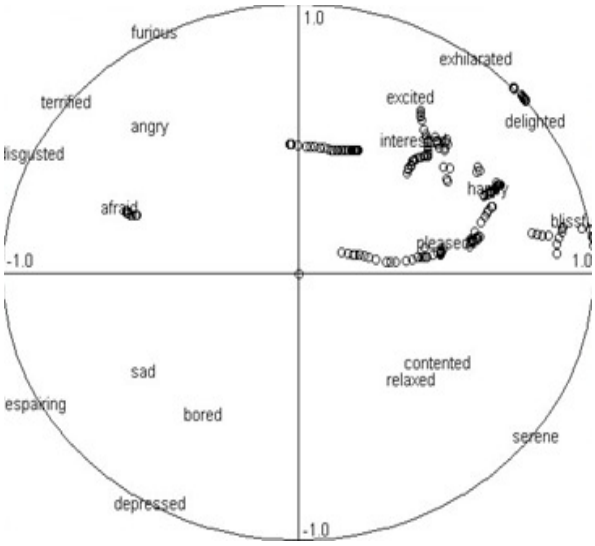

Figure 5. When mapped against Russell's Circumplex Model, a few laughter instances with label hurtful were found on the positive arousal, negative valence quadrant.

6. Conclusion

Based on our findings we see that indeed, humans tend to use laughter to express various emotions such as happiness, hurtful, giddiness, embarrassment and excitement. To label correctly, annotators had to rely not only on multimodal information of face and voice, but also on contextual information. Using manual observation, it appears that the voice is a better modality to distinguish between the different kinds of laughter.

7. Acknowledgements

The authors would like to thank De La Salle University's University Research and Coordination Office (URCO) for the funds used for this project, including all the human subjects and the annotators.

We especially would like to thank Julie Ann Alonzo, Janelle Marie Campita, Stephanie Therese Lucila, Miguel Crisanto Miranda, Christopher Galvan, David Manangan, Michael Sanchez and Jason Wong for their invaluable contribution to this work.

8. References

Bahon-Cohen, S. & Wheelwright, S. (2004). *The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Difference.* Journal of Autism and Developmental Disorders. Vol. 34, No. 2, pp. 163 – 175.

Boersma, P. and Weenink, W. (2009). PRAAT: Doing Phonetics by Computer Version 5.1.05 [Computer program], Retrieved June 2009, from http://www.praat.org.

Chellapa, R., & Zhao, W. (2006). *Face Processing*. Burlington, MA: Elsevier/Academic Press.

Cowie R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. (2000). *Feeltrace: An Instrument for Record-ing Perceived Emotion in Real Time*. Proc. ISCA Workshop Speech and Emotion, pp. 19-24, 2000.

Devillers, L., & Vidrascu L. (2007). *Positive and Negative emotional states behind the laughs in spontaneous spoken dialogs*. Interdisciplinary Workshop on The Phonetics of Laughter. 37-40.

Escalera, S., Puertas, E., Oriol, P. and Radeva, P. 2009. Multi-modal Laughter Recognition in Video Conversations. *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference*.

Knox, M., Morgan, N., and Mirghafori, N. 2008. Getting the last laugh: Automatic laughter segmentation in meetings. *Interspeech 2008*.

Masuda, T.; Ellsworth, P.; Mesquita, B. Leu, J.; Tanida, S. & Van der Veerdonk, E. (2008). Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology*, 94: 365-381

Matsumoto, D., Yoo, S-H., Fontaine, J., Anguas-Wong, A. M., Arriola, M.,Ataca, B., et al. (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism vs. collectivism. *Journal of Cross-Cultural Psychology, 39,* 55–74.

Mesquita, B. & Frijda, N.H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin*, 112: 179-204.

Nachamai, M., & Santhanan, T. (2008). *Laughter inquisition in affect recognition*. Journal of Theoretical and Applied Information Technology. 429-432.

Pantic, M. and Petridis, S. (2010). *Classifying Laughter and Speech using Audio-Visual Feature Prediction*. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.

Pantic, M., & Petridis, S. (2008). *Audiovisual discrimination between laughter and speech.* Acoustics, Speech and Signal Processing. 5117-5120.

Petridis, S.; Asghar, A. & Pantic, M. (2010). Classifying laughter and speech using audio-visual feature prediction, in *IEEE ICASSP*, pp. 5254–5257

Reuderink B. , Poel M., Truong K., Poppe R., Pantic M. (2008). *Decision-level fusion for audio-visual laughter detection*. MLMI '08: Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction, Springer-Verlag, Berlin, Heidelberg, pp. 137–148.

Russel J. (1980). *A Circumplex Model of Affect.* Journal of Personality and Social Psychology. Vol. 39. No. 6. pp. 1161 – 1178. Truong, K. P. and van Leeuwen, D. A. (2007). *Automatic discrimination between laughter and speech*. Speech Communication. Volume 49.

Safdar, S., Friedlmeier, W., Matsumoto, D.; Yoo S-H., Kwantes, C., Kakai, H. et al. (2009). Variations of emotional display rules within and across cultures: A comparison between Canada, USA, and Japan. *Canadian Journal of Behavioural Science*, 41: 1-10.

Scherer, S.; Schwenker, F.; Campbell, N & Palm, G. (2009). Multimodal laughter detection in natural discourses, *Human Centered Robot Systems*, pp. 111–120.

Truong, K. P., Van Leeuwen, D. A. (2005) Automatic detection of laughter. Proc. of Interspeech. pp. 485-488.

Truong, K. P. & Van Leeuwen, D. A. (2007) Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features. Workshop on the Phonetics of Laughter. Saarbrucken.