Prediction of Shared Laughter for Human-Robot Dialogue

Divesh Lala
Kyoto University
Japan
lala@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Kyoto University
Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Kyoto University
Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

Shared laughter is a phenomenon in face-to-face human dialogue which increases engagement and rapport, and so should be considered for conversation robots and agents. Our aim is to create a model of shared laughter generation for conversational robots. As part of this system, we train models which predict if shared laughter will occur, given that the user has laughed. Models trained using combinations of acoustic, prosodic features and laughter type were compared with online versions considered to better quantify their performance in a real system. We find that these models perform better than the random chance, with the multimodal combination of acoustic and prosodic features performing the best.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

shared laughter; machine learning; conversation; human-robot dialogue

ACM Reference Format:
Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2020. Prediction of Shared Laughter for Human-Robot Dialogue. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425265

1 INTRODUCTION

Although robots and agents are becoming more human-like in appearance and movement, their ability to engage in free conversation is still at the surface level. This not only applies to sophisticated dialogue generation through natural language processing, but other non-linguistic phenomena such as backchanneling and fillers which are a part of natural conversation.

One aspect of these systems is that of laughter during conversation. In particular, spontaneous laughter should be a goal of such systems as this is not only human-like, it allows for more engagement with the user [20, 24]. There have been attempts to create systems which use linguistics to recognize humor [4, 6, 22], but modeling these in live interactions is still difficult due to speech recognition errors and added complexity of face-to-face interaction. Another method is to recognize a laugh from the user and join in with them, using behavior from the user as an indication of when a laugh is appropriate. This can be termed as shared laughter.

The ongoing long-term goal of this research is to implement a shared laughter system for the female android ERICA [10]. We propose that such a system will add to the naturalness of ERICA in free conversation. We also propose that such a system is not a simple process of recognizing the user's laugh and joining in, but involves several classification steps.

Shared laughter has been implemented in systems with an external stimulus, such as a video [20], so the system can be sure that laughing with the user is suitable. On the other hand, in face-to-face conversation there is no external stimulus and less certainty about whether joining in with the user is acceptable. In real conversation many laughs are self laughs with no response from the conversation partner. Therefore knowing if the system should engage in shared laughter is important to maintain naturalness in conversation.

Our main goal in this paper is to address the challenge of distinguishing between self laughs and shared laughs. This model can then be implemented in a conversational system to exhibit natural laughing behavior. We measure the performance of several types of models which use acoustic, prosodic features and laughter type. Furthermore, we perform online model evaluation to better quantify model performance for a real system.

2 RELATED WORK

Laughter detection from audio and visual sources has been well studied [1, 2, 7, 11, 18, 23, 25]. In this work we assume inital laughter detection can be achieved using an external model and so focus on further classification of shared laughter. Shared laughter itself has been proposed as a form of mimicry [9, 19], studied in terms of the relation of each laugh's intensity [8], and is associated with different speaker behaviors compared to self laughter [12].

There has also been research on laughter for robots and agents including linguistic studies [3, 5], laughter motion generation [17] and shared laughter responses [26]. Integrating laughter in a robot or agent has also been shown to increase engagement and amusement [20, 24]. Studies on the application of robot laughter have so far been confined mainly to scenarios where an external stimulus provides the trigger for laughter. For example, a video [13, 20] or multi-party quiz [24]. On the other hand, our aim is for shared laughter to occur within some conversation, where the trigger is laughter through dialogue rather than an external stimulus.

We are unaware of any models which have been trained on this specific task. Exploring this aspect of conversation is necessary

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425265

since we want the robot to engage in shared laughter at appropriate times, not every time a laugh is detected from the user.

3 SHARED LAUGHTER MODEL

We propose a general system model of shared laughter consisting of three main modules, as shown in Figure 1. We assume that shared laughter consists of the user's *initial laugh*, followed by the system's *response laugh*.

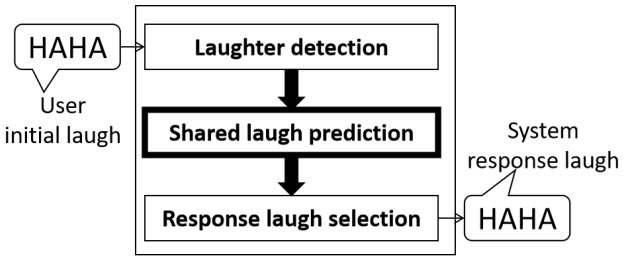

Figure 1: General overview of shared laughter system. The focus of this work is the shared laughter prediction module.

The first step is the detection of the initial laugh. This paper assumes we have a model which can predict if an inter-pausal unit (IPU) contains laughter or not. In Section 6.2 we show how much this affects the performance of the entire system.

The second module is our main focus, which decides if a response laugh should occur to create shared laughter. A system which always responds with a laugh is far from human-like, as we will show in Section 4. Furthermore, there are situations where it may be undesirable to respond with laughter, such as embarrassed laughter from the user. We propose that a shared laughter model should predict suitable instances where laughter can be shared.

The final module is the decision of what type of laugh the agent should respond with. We do not focus on this function in this work, but propose that it is important for the agent to use several laughs and pick one which is of an appropriate tone and emotion.

4 DATA COLLECTION AND ANALYSIS

We use data from a corpus of human-robot dialogues with a tele-operated ERICA, collected from interactions with 61 male participants where ERICA was controlled by one of four female Wizard of Oz operators. The scenario was that of speed dating - participants engage in casual chat about topics such as favorite hobbies, likes and dislikes. Operators engaged in completely free chat, allowing explicit spontaneous laughs from operators and participants. Although it is clear the "'date"' with ERICA is not real, this scenario is an ideal example of mixed initiative conversation.

Each IPU was transcribed and laughs were annotated into two types. A laugh occurring by itself was considered an *isolated laugh*, while a laugh which occurred as part of speaking was considered a *speech laugh*. We consider any IPU which is either an isolated laugh or a speech laugh to be under the general category of a *laugh IPU*.

Given an initial laugh, we consider it to be shared if all the following conditions hold:

- the initial laugh IPU is part of the speaker's turn
- the initial laugh IPU is an isolated laugh or a speech laugh which ends with laughter
- the response laugh occurs within two seconds after the initial laugh has ended
- both the initial laugh IPU and response laugh are greater than 400ms in length

These conditions are somewhat arbitrary but the final condition is based on ERICA's actual system that ignores very short IPUs. Laughs which matched these criteria were considered to be shared laughs. Laughs which were classified as response laughs are not considered to be initial laughs and so are omitted from the analysis. All other laughs are considered to be self laughs. We identified 1206 initial laughs which fit these criteria, 698 (57.9%) self laughs and 508 (42.1%) shared laughs.

Table 1 shows statistics for laughter type according to the speaker. Subjects laughed much more than operators. The majority of subjects laughs were self laughs, while the majority of operator laughs were shared. This could be due to the nature of the scenario, where subjects may have been more proactive during speed dating.

Table 1: Frequencies of laughter type vs. speaker

	Operator	Subject	Total
Self	58	640	698
Shared	232	276	508
Total	290	916	1206

Table 2 shows statistics for laughter type according to how it was generated. Speech laughs outnumbered isolated laughs and shared laughter is more likely to have a speech laugh as the trigger (73%).

Table 2: Frequencies of laughter type vs. generation

	Isolated	Speech laugh	Total
Self	248	450	698
Shared	139	369	508
Total	387	819	1206

This data suggests that most laughs are not shared, so a conversational system should make a decision of which initial laughs are appropriate for a response laugh. It also shows that laughter detection should identify both isolated and speech laughs.

We use these laugh IPUs as samples for our models. Filterbanks, pitch and power were extracted at 100Hz using an online pitch tracker [15], with a microphone array for the subject and shotgun microphone for the operator. The microphone array environment is the same as the one used in ERICA's live system [16], so we can accurately determine how well the model would work in real time.

5 MODEL CREATION

We now describe several models which will make a classification decision on whether a laugh is a self laugh or a shared laugh. Given

that a laugh IPU has been recognized, can we predict if a response laugh is uttered?

First we consider the issue of how to deal with the properties of isolated and speech laughs. For an isolated laugh the entire IPU will contain laughter information, but for a speech laugh the relevant information is at the end of the IPU. From preliminary analysis, we hypothesized that using just the final 1000ms of audio (whether it is an isolated laugh or a speech laugh) could perform as well as using the entire audio as a sample. In Section 6 we compare using the final 1000ms (100 frames) of audio of the laugh IPU as a sample to the entire laugh IPU.

Two types of audio-based features were considered as inputs to the model. The first are the means and standard deviations of 40 acoustic mel-filterbank features. We chose these features for practical reasons, since they can be extracted in ERICA's real time system and are also used for our laughter detection system. Means and standard deviations of acoustic features have also been used in previous work on laughter [26].

The second type of features are prosodic. These are based on the pitch and power values across the entire IPU which can be easily calculated in real-time. Specifically, for both pitch and power we calculate the mean, median, standard deviation, maximum, minimum and range. For pitch values, we only consider frames which are voiced, since including unvoiced frames would have a significant effect on the statistics. We also include the proportion of frames in the IPU which are voiced as a separate feature, and the total duration of the laugh, for a total of 14 prosodic features.

We also include the laughter type (isolated or speech laugh) as a binary feature. From our corpus analysis, we expect that shared laughter will be positively associated with a speech laugh.

We also explored linguistic features by using word vectors of the previous utterances leading up to a laugh, and facial features extracted by a web camera. However these performed worse than the audio features, so have been omitted. This could be a result of the comparatively low number of samples in our corpus.

All features were standardized for model training and feature comparison. We trained logistic regression (LR) and support vector machine (SVM) models. We also attempted deep learning techniques but these showed poor performance, perhaps due to the low number of samples. Training was implemented using 10-fold cross-validation. Furthermore, we only consider initial laughs from male users as part of the samples. This was because ERICA was controlled by one of only four female operators with similar voices. Therefore the data set contains 916 samples, 640 self laughs (69.9%) and 276 shared laughs (30.1%).

6 RESULTS

Results of the models are given below, measuring the performance on the prediction of shared laughter. Performance can be measured in two ways - offline and online. The differences are shown in Figure 2. In the offline version, we assume that laughter detection and laugh type detection (if any) is perfect. In the online version, we have to make assumptions about errors in the initial laughter detection model and the laughter type which will degrade the performance of the model, but give us a more accurate indication of the actual performance of the model in a live system.

The online evaluation will assume that ERICA's current laughter detection model [14] will predict if the initial IPU is a laugh or not. It was trained on the same interaction corpus and has an F-score of 0.762. We only test initial laughs where a laugh actually happened, since our goal is shared laughter. If the laughter detection model predicts a non-laugh, the shared laugh model will automatically output no response laugh. For models with laughter type as a feature, we use another classification model which has an F-score of 0.905 to predict if a laugh is isolated or a speech laugh, and use this result in the shared laughter model.

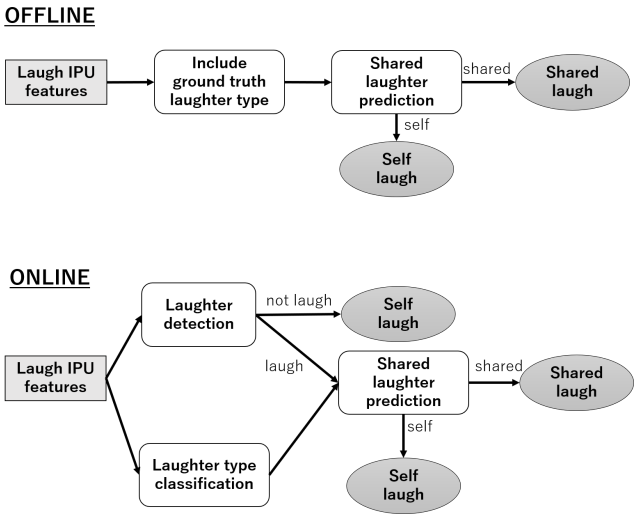

Figure 2: Diagrams showing the prediction process of offline (top) and online (bottom) models.

The baseline is a random model which selects the class according to the distribution of shared laughter in the corpus (30.1%). We use this as a baseline since our corpus shows that a model which always does shared laughter would be unsuitable in a live system.

6.1 Offline evaluation

We first present results of the offline system. The laugh IPU is assumed to have already been correctly classified as a laugh and as a speech-laugh or isolated laugh.

We evaluate models based on how much of the audio sample could be used for prediction - the whole IPU or only the last 1000ms (if the IPU is longer than 1000ms). This is because some samples are long speech-laugh IPUs. By including only the end of the IPU we hope to remove redundant information not related to the acoustic properties of the laugh. Table 3 presents results of our models. For brevity, we only display the result of the best type of model (logistic regression or SVM) for each combination of features.

We find that using the last 1000ms of a laugh IPU results in performance comparable to using the entire IPU, although the best performing model combined acoustic features and the laugh type, while using all features results in a weaker model. Using laughter type as a feature tends to improve the recall of the model.

Table 3: Performance comparison of offline models

	Precision	Recall	F-score
Baseline	0.301	0.301	0.301
Laugh type	0.334	**0.743**	0.461
Whole laugh IPU			
Acoustic (SVM)	0.396	0.601	0.477
Prosodic (LR)	0.375	0.449	0.409
Aco. + pros. (SVM)	0.404	0.612	0.487
Pros. + laugh type (SVM)	0.319	0.670	0.432
Aco. + laugh type (SVM)	0.415	0.645	**0.505**
All features (SVM)	0.406	0.616	0.489
Last 1000ms of IPU			
Acoustic (SVM)	0.421	0.591	0.492
Prosodic (SVM)	0.369	0.453	0.407
Pros. + laugh type (SVM)	0.333	0.707	0.453
Aco. + pros. (SVM)	0.415	0.601	0.491
Aco. + laugh type (SVM)	0.421	0.620	0.501
All features (SVM)	**0.422**	0.594	0.493

Through feature selection across all folds, we found the prosodic model selected just three features which had consistently high absolute values for their coefficients, which could be directly compared since they were all previously standardized. These were the mean pitch (-0.44), median pitch (0.52) and length of the laugh IPU (0.37). This suggests that a laugh is more likely to be shared if it has a pitch distribution with a negative skew (since the median pitch is much greater than the mean pitch) and lasts longer than average.

6.2 Online evaluation

Results of the online model's performance are shown in Table 4, where we consider the actual output of laughter detection and laugh type classification models. We compare the best offline models (according to F-score) to the online version.

Table 4: Performance comparison of online models

	Prec.	Rec.	F-score
Baseline	0.301	0.301	0.301
Acoustic offline	0.421	0.591	0.492
Acoustic online	0.412	0.464	0.440
Prosodic offline	0.375	0.449	0.409
Prosodic online	0.341	0.558	0.423
Aco. + pros. offline	0.415	0.601	0.491
Aco. + pros. online	0.413	0.489	0.448
Prosodic + laugh type offline	0.333	0.707	0.453
Prosodic + laugh type online	0.327	0.536	0.407
Acoustic + laugh type offline	0.415	0.645	0.505
Acoustic + laugh type online	0.413	0.478	0.443
All features offline	0.422	0.594	0.493
All features online	0.412	0.489	0.447

Results show a clear degradation in performance due to errors from laughter detection and laugh type classification, although all models still perform above the baseline. Recall is more affected in an online setting, which is intuitive since the laughter detection model will produce false negatives. The best online model uses both acoustic and prosodic features.

7 DISCUSSION

Our results show that the acoustic and prosodic features are able to detect whether a laugh is a self laugh or a shared laugh better than chance, although the performance of the model is still fairly weak. The analysis of prosodic features suggested the acoustics of the initial laugh may have some characteristics which encourage a response laugh. We also showed that the performance of laughter detection is influential on the overall performance of the system so this must also be improved for better classification performance.

We only trained on male subjects since the nature of the experiment meant that operator laughs could only be elicited from a small number of women. Gender could have an effect on laughter, particularly for a speed dating scenario, so more female laughter samples should produce a more generalizable model.

To progress towards our final system, we should take advantage of other features and deep learning techniques. Due to the small sample size in this study, these were not effective in this work, but we expect other modalities become useful if we have additional data, particularly facial expressions, linguistics and sentiment [21] to play a role as an indication of emotion and possible humor.

An obvious issue is whether in fact we can consider the corpus to be a "ground truth". There are many instances where a shared laugh is appropriate but not executed or vice versa. Another approach to address this is manual annotation of only "strong" and "weak" shared laughter, although this would reduce the number of training samples. We intend to conduct a subjective experiment with our final shared laughter system to evaluate if our approach is useful. These results would give us a better indication of the quality of our system than classification performance.

One aspect this study lacks is the timing of shared laughter. The live version of our system assumes that the entire laugh has been segmented before the robot can respond to it. However in the corpus there are many examples of overlapping shared laughter. Although it is possible to segment the user's laughs quickly and respond straight after, we do not know if this is acceptable compared to overlapping laughter. This can be addressed in future work.

8 CONCLUSION

In this work we developed models for predicting shared laughter in human-robot dialogue. We found that combining acoustic and prosodic features was the best performing in an online system, although there is much room for improvement, and initial laughter detection has a significant impact on model performance. We intend to implement this in a robot as part of an overall shared laughter system for subjective evaluation.

ACKNOWLEDGMENTS

This work was supported by JST ERATO Grant number JPMJER1401 and JSPS KAKENHI Grant number JP19H05691.

REFERENCES

[1] Zahid Akhtar, Stefany Bedoya, and Tiago H Falk. 2018. Improved Audio-Visual Laughter Detection Via Multi-Scale Multi-Resolution Image Texture Features and Classifier Fusion. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3106–3110.

[2] Faramarz Ataollahi and Merlin Teodosia Suarez. 2019. Laughter Classification Using 3D Convolutional Neural Networks. In *International Conference on Advances in Artificial Intelligence (ICAAI)*. 47–51.

[3] Anton Batliner, Stefan Steidl, Florian Eyben, and Björn Schuller. 2019. On Laughter and Speech-Laugh, Based on Observations of Child-Robot Interaction. *arXiv preprint arXiv:1908.11593* (2019).

[4] Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *North American Chapter of the Association for Computational Linguistics (NAACL)*. 130–135.

[5] Francesca Bonin, Nick Campbell, and Carl Vogel. 2014. Time for laughter. *Knowledge-Based Systems* 71 (2014), 15 – 24. https://doi.org/10.1016/j.knosys.2014.04.031

[6] Lei Chen and Chong MIn Lee. 2017. Convolutional neural network for humor recognition. *arXiv preprint arXiv:1702.02584* (2017).

[7] Sarah Cosentino, Salvatore Sessa, and Atsuo Takanishi. 2016. Quantitative laughter detection, measurement, and classification—A critical survey. *IEEE Reviews in Biomedical engineering* 9 (2016), 148–162.

[8] Kevin El Haddad, Sandeep Nallan Chakravarthula, and James Kennedy. 2019. Smile and Laugh Dynamics in Naturalistic Dyadic Interactions: Intensity Levels, Sequences and Roles. In *International Conference on Multimodal Interaction (ICMI)*. 259–263.

[9] Sarah Estow, Jeremy P Jamieson, and Jennifer R Yates. 2007. Self-monitoring and mimicry of positive and negative social behaviors. *Journal of Research in Personality* 41, 2 (2007), 425–433.

[10] Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. Erica: The ERATO intelligent conversational android. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 22–29.

[11] Gábor Gosztolya and László Tóth. 2019. Calibrating DNN Posterior Probability Estimates of HMM/DNN Models to Improve Social Signal Detection From Audio Data. In *Interspeech*. 515–519.

[12] Rahul Gupta, Theodora Chaspari, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. Analysis and modeling of the role of laughter in motivational interviewing based psychotherapy conversations. In *Interspeech*.

[13] Jennifer Hofmann, Tracey Platt, Willibald Ruch, Radoslaw Niewiadomski, and Jérôme Urbain. 2015. The influence of a virtual companion on amusement when watching funny films. *Motivation and Emotion* 39, 3 (2015), 434–447.

[14] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. *APSIPA Transactions on Signal and Information Processing* 7 (2018).

[15] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50, 6 (2008), 531–543.

[16] Carlos T. Ishi, Chaoran Liu, Jani Even, and Norihiro Hagita. 2016. Hearing support system using environment sensor network. In *IROS*. 1275–1280.

[17] Carlos Toshinori Ishi, Takashi Minato, and Hiroshi Ishiguro. 2019. Analysis and generation of laughter motions, and evaluation in an android robot. *APSIPA Transactions on Signal and Information Processing* 8 (2019).

[18] Reshmashree B Kantharaju, Fabien Ringeval, and Laurent Besacier. 2018. Automatic Recognition of Affective Laughter in Spontaneous Dyadic Interactions from Audiovisual Signals. In *International Conference on Multimodal Interaction (ICMI)*. 220–228.

[19] Costanza Navarretta. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *International Conference on Language Resources and Evaluation (LREC)*. 469–474.

[20] Radosław Niewiadomski, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Huseyin Cakmak, Sathish Pammi, Tobias Baur, Stephane Dupont, Matthieu Geist, Florian Lingenfelser, Gray McKeown, Olivier Pietquin, and Willibald Ruch. 2013. Laugh-aware virtual agent and its impact on user amusement. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 619–626.

[21] Birgitta Ojamaa, Kristiina Jokinen, and Kadri Muischenk. 2015. Sentiment analysis on conversational texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. 233–237.

[22] Anil Ramakrishna, Timothy Greer, David C Atkins, and Shrikanth Narayanan. 2018. Computational Modeling of Conversational Humor in Psychotherapy. In *Interspeech*. 2344–2348.

[23] Khiet P Truong and David A Van Leeuwen. 2007. Automatic discrimination between laughter and speech. *Speech Communication* 49, 2 (2007), 144–158.

[24] Bekir Berker Türker, Zana Buçinca, Engin Erzin, Yücel Yemez, and T Metin Sezgin. 2017. Analysis of Engagement and User Experience with a Laughter Responsive Social Robot. In *Interspeech*. 844–848.

[25] Bekir Berker Turker, Yucel Yemez, T Metin Sezgin, and Engin Erzin. 2017. Audio-facial laughter detection in naturalistic dyadic conversations. *IEEE Transactions on Affective Computing* 8, 4 (2017), 534–545.

[26] Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. 2009. AVLaughterCycle: An audiovisual laughing machine. In *International Summer Workshop on Multimodal Interfaces*. 79–87.