An evaluation of manual and semi-automatic laughter annotation

Bogdan Ludusan, Petra Wagner

Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC,
Bielefeld University, Germany

{bogdan.ludusan,petra.wagner}@uni-bielefeld.de

Abstract

With laughter research seeing a development in recent years, there is also an increased need in materials having laughter annotations. We examine in this study how one can leverage existing spontaneous speech resources to this goal. We first analyze the process of manual laughter annotation in corpora, by establishing two important parameters of the process: the amount of time required and its inter-rater reliability. Next, we propose a novel semi-automatic tool for laughter annotation, based on a signal-based representation of speech rhythm. We test both annotation approaches on the same recordings, containing German dyadic spontaneous interactions, and employing a larger pool of annotators than previously done. We then compare and discuss the obtained results based on the two aforementioned parameters, highlighting the benefits and costs associated to each approach.

Index Terms: laughter, speech-laugh, annotation, manual, semi-automatic

1. Introduction

Laughter is a pervasive phenomenon in human communication, fulfilling various roles and functions in spoken interaction [1]. Although it has seen an increased interest, both from a scientific perspective [1] and a technological viewpoint [2], there have been very few cross-linguistic laughter studies. To our knowledge, the acoustic characteristics of laughter [3], laughter position compared to its laughable [4], laughter entrainment [5] and the distribution of laughter across linguistic levels [6] are the only laughter aspects investigated across languages.

One of the reasons for this apparent lack of cross-linguistic laughter studies seems to be the scarcity of high-quality multilingual resources annotated for laughter (one notable exception being the DUEL corpus [7]). Although datasets including laughter annotation exist in several languages, the type of materials contained might not be similar enough for a cross-linguistic comparison, or the differences in the annotation guidelines employed might not allow it. Moreover, with laughter being a secondary goal of the labelling process, issues with the resulting annotation may be observed [8]. A possible solution to this problem lies in employing already existing comparable materials in different languages. In this study, we take a first step in this direction, by (1) evaluating the process of annotating laughter in existing corpora and (2) investigating ways to facilitate this process.

On the one hand, we analyze the time needed to manually annotate laughter in spontaneous interactions and the inter-rater reliability that can be observed for this process. To our knowledge, there has been no study reporting annotation times for laughter. A better understanding of the required time would allow annotation campaigns to be planned more realistically. With regards to inter-rater reliability, the majority of studies have looked at specific laughter sub-types or characteristics, such as: speech-laughs [9], laughter voicedness [10, 11] or laughter production type, style, function and its vowel quality [12]. Only recently, Truong and colleagues [13] have performed an evaluation of the overall laughter annotation process, for a number of corpora and across the different levels of a laughter episode. Although they do not compute inter-rater reliability, they report the percentage of matching units found between annotators, which is, in itself, a measure of annotators' agreement. In our study, we recruited a larger number of participants in an annotation experiment and calculated standard inter-rater reliability measures such as κ. Differently from [13], who employed annotators with a very high level of expertise, our annotators had a more limited experience with laughter annotation, similar to real-life conditions in corpus building. Thus, we hope to obtain a more ecological account of inter-annotator agreement.

On the other hand, we investigate how one could automate the annotation process, as well as to comprehend the associated benefits and costs. While automatic laughter detection systems have been previously proposed (e.g. see the 2013 Paralinguistic Challenge [14]), we focus here on a semi-automatic process, in order to be able to use its output also in more fine-grained analyses. [15] has put forward a semi-automatic approach for laughter annotation, involving a first step in which a number of laughter events were manually annotated and then used to train classification models based on audio-visual features. Laughter events were predicted based on the trained models and their boundaries manually refined in a subsequent step. Here, we take a different approach: rather than positing laughter event boundaries and then manually correcting them, we build a system that returns the speech intervals that might contain laughter, followed by manual placing of the boundaries of laughter events occurring in these intervals. For this, we employed rhythm information proved to be useful for the perception of laughter [16]. In particular, we utilized a representation called the modulation index spectrum [17], which has been shown to discriminate between laughter and speech [18]. We developed an automatic tool that exploits this information and returns the segments of time where laughter is likely to occur. The annotators were then asked to focus their attention on these intervals and to manually mark the boundaries of perceived laughter events. We compare this semi-automatic process with the manual one, based on the required annotation time and on the inter-rater reliability.

2. Dataset

We performed our investigation on the DUEL corpus [7], a dataset containing spontaneous interactions in three languages: French, German and Mandarin Chinese. The corpus was annotated at the utterance level and includes conversational speech phenomena such as laughter (both laughs and speech-laughs – simultaneous occurrences of speech and laughter, in which nei-

ther of the two components is dominant).

We employed German recordings from two scenarios: Dream Apartment and Film Script. In the first one, the participants discussed the design, furnishing and decoration of an apartment they would share. In the second one, the interlocutors were requested to come up with a film script based on an embarrassing moment. For the development of the tool, we used the same materials as in [18], approximately two hours of data, corresponding to the recordings of eight dyads from the Film Script scenario. Seven out of the eight pairs consisted of friends/colleagues (11 females and 5 males). Based on the eight recordings, representations for the speech, laughs and speech-laugh classes were computed. One of the recordings (15 minutes, 2 females) was further used to set a number of parameters employed by the tool for deciding whether laughter is present in the speech signal. The tool was tested on one hour of data, belonging to four pairs of speakers (6 females, 2 males) discussing the Dream Apartment scenario. The pairs were different from those whose data was used for the development of the tool and were chosen because they produced the most laughter among the dyads of that scenario. One of the four recordings (15 minutes) was further employed in the annotation experiment.

3. Methods

3.1. Automatic laughter pre-annotation tool

The proposed laughter annotation tool was built on a signal-based description, the modulation index spectrum, shown to be able to capture aspects related to speech rhythm [17]. Moreover, it can reliably discriminate between laughs and speech and between speech-laughs and speech [18]. Here, we employ the same materials as in the aforementioned study, in order to derive a representation that maximizes the difference between the laugh and speech-laugh classes, on the one hand, and the speech class, on the other.

First, we computed the modulation index, as in [18], for each of the three classes of vocalizations (laugh, speech-laugh, speech), by means of the AM_FM_Spectra toolbox [19]. The input speech signal was band-pass filtered with a 30-channel gammatone filterbank, each channel being 1 ERB-wide and having their center frequencies equally spaced on the ERB scale. The responses of the filters had their envelopes extracted and were further filtered using a bank of $1/3$ octave wide Butterworth bandpass filters. In a last step, the root-mean-square amplitude of each filter output was divided by the mean amplitude of the output of the gammatone filter. To obtain the spectrogram-like description (a 2D matrix), we stacked the output of all 30 channels, having the modulation rate on the horizontal axis, the audio frequency on the vertical axis and the modulation index amplitude as the value of each point.

Next, we took the matrices corresponding to the laugh and speech-laugh classes, computed their average and subtracted the speech class matrix. The resulting *difference matrix*, illustrated in Figure 1, informs the automatic tool which modulation rate/audio frequency points to consider when computing a laughter detection function, by applying a threshold ($quanThr$) based on its quantile values.

Then, at run time, we computed a detector function based on this information. Since rhythm is a suprasegmental phenomenon, and following the same settings as in [18], we used an analysis frame of 1.5 seconds (s), with a 50 ms overlap. For each analysis frame, we took only the points of its modulation index spectrogram which surpassed the $quanThr$ parameter value in

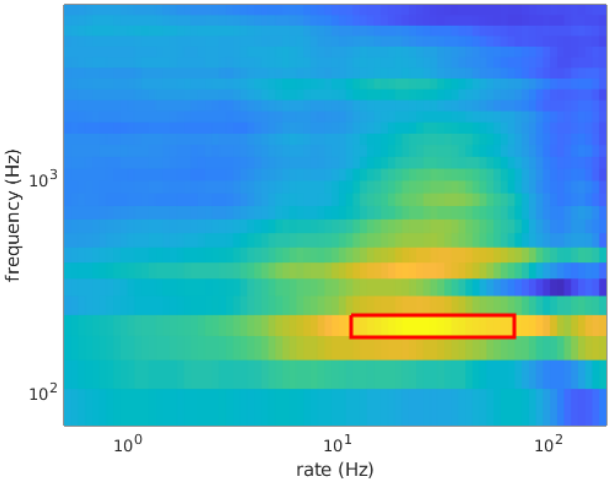

Figure 1: *Modulation index spectrogram difference between the average of the laugh and speech-laugh classes, on the one side, and that of the speech class, on the other. The area delimited by a red rectangle represents the points considered for the computation of the detector function, based on the parameter value ($quanThr = .99$) established on the dev set.*

the *difference matrix* (see Figure 1). The mean value of these points represents the detector function value for that particular frame. The resulting detector function (for an example, see Figure 2, upper panel) was then smoothed using a sliding window of length $movWin$ frames, centered on the current sample.

Finally, we determined candidates for intervals containing laughter events by searching for the local maxima of the detector function. The candidates were restricted to those having a minimum peak prominence, compared to the neighbouring valleys, of at least $minProm$ (cf. Figure 2, middle panel).

We evaluated the goodness of the tool by comparing the output obtained on the test set to the reference laughter annotation supplied with the corpus. This was done by means of standard measures: precision (how many laughter events, out of the total number of events found, were correct), recall (how many events, out of the total number of events in the reference data, were found) and F-score (the harmonic mean of precision and recall). We considered a candidate to be correct if a reference event was found within a 5 s window centered at the time of the local maximum of the detector function (marked as int in Figure 2, lower panel). We decided to consider such a large window around the maximum, as the system is based on a large analysis window (1.5 s) and several consecutive laughter events might be represented by only one peak in the detector function. This decision was also consistent with the second part of the annotation process, in which the annotators had to manually place the boundaries of a laughter event, by listening to a 5 s speech segment centered on each candidate.

The three parameters employed by the algorithm ($quanThr$, $movWin$ and $minProm$) were determined on a small development set consisting of the recordings of one dyad, by finding the triple that maximized F-score, provided that precision was higher than 0.8. This condition was set in order to obtain an algorithm that retrieves a high number of laughter events, while keeping the number of false positives low. The right compromise between recall and precision had to be found, as a high recall but a low precision would defeat the purpose

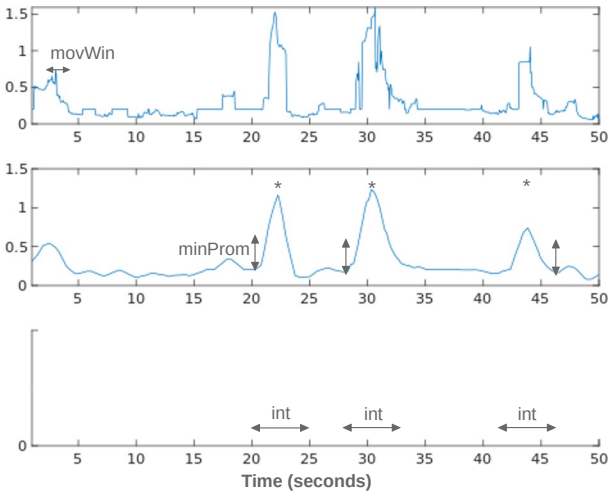

Figure 2: *Steps followed in the automatic tool: upper panel – smooth the detector function with a moving window, middle panel – find laughter event candidates by applying a minimum prominence threshold to the function, and lower panel – return for manual annotation intervals centered on the candidates.*

of the tool, namely reducing the amount of recordings that the annotators would have to listen to in order to label laughter events. Conversely, a very high precision but a low recall, would render the tool useless, as it would flag only a small number of laughter candidates to be checked by the annotators.

Since existing corpora are, in general, annotated for voice activity detection, we also used this information in our tool: no modulation index was computed for the analysis frames which were centered on a frame marked as being silence in the reference transcription. At run time, the frames corresponding to silence were replaced by the median value of the detector function, before the moving average smoothing was applied.

3.2. Laughter annotation experiment

An experiment was set-up in order to evaluate the manual and the semi-automatic annotation processes. Seven phonetically-trained annotators received the recordings corresponding to one of the dyads discussing the Dream Apartment scenario. The criterion behind the choice of the dyad was the similarity of the per-dyad automatic tool performance to the overall test set results. The recordings were supplied with each speaker (channel) as a separate file, for an easier labelling process in the case of overlapping laughter between the speakers.

The annotators were given a Praat [20] TextGrid file with two tiers, the first tier containing the intervals they were supposed to listen to for the experiment. As the tool employed in the semi-automatic process had knowledge of the stretches of speech/silence in the recordings, we made available this information for both annotation tasks. Thus, for the manual task, the first tier contained the regions of speech in the recording, as in the reference. For the semi-automatic task, the first tier consisted of 5 s long intervals, centered at the points in time corresponding to the candidates returned by the tool. In case two such intervals overlapped (the distance between two laughter candidates was lower than 5 s), they were merged.

The participants were asked to mark the boundaries of the perceived laughter events on the second tier of the TextGrid file. They annotated two types of laughter events: laughs and speech-laughs, by labelling them with the labels L and SL, respectively. They were allowed to take breaks during the experiment, provided that the duration of the break was subtracted from the total duration of the annotation process. All annotators performed both tasks, in order to be able to directly compare results. We balanced the order in which the tasks were done, with three participants carrying out the manual process first, while the remaining four completed the semi-automatic process first. Moreover, to reduce the effect of the first task on the second one, they were performed in different weeks, 5 to 6 days apart.

Table 1: *Automatic tool performance on the dev set, test set and the recording used for the annotation experiment.*

Set	Precision	Recall	F-score
dev	.807	.698	.749
test	.565	.891	.691
experiment	.606	.891	.722

Each task of the experiment had two outcomes: the time (in minutes) required for the participants to complete it and the TextGrid file containing the laughter event boundaries. The latter was further used to calculate two measures of inter-rater reliability, the overall agreement and the κ measure. The overall agreement was computed as the proportion of annotator-pairs agreeing, out of the total number of annotator-pairs, averaged across all observations. The kappa measure is a reliability metric which normalizes the overall agreement by the probability of chance agreement by the annotators. Here, we chose a free-marginal multirater κ [21] since the performed annotation process has no a-priori number of cases that should be assigned to each category. All inter-rater measures were computed using an online tool [22]. We also determined the number of missed laughter events, defined as being the events found after the manual task, but which no annotator marked as such in the semi-automatic process. For calculating the reliability we considered two annotated events to be corresponding if there was a minimum overlap of 50% between them. Consecutive events with the same label and a pause shorter than 50 ms between them counted as one event.

4. Results

The results attained on the development and the test set, using the parameters determined on the dev set ($quanThr = .99$, $movWin = 9$ frames and $minProm = 0.27$), are presented in Table 1. We then selected the test recording which had the most similar performance to the overall test set (cf. last line of Table 1) and used it in the annotation experiment.

The time required by the annotators to complete the manual task was, on average, 75 minutes (stdev: 25), while the mean time needed for the semi-automatic task was 52 minutes (stdev: 17). The differences were found significant ($p = .031$) when tested with a Wilcoxon signed-rank test. Per-speaker details are presented in the second column of Table 2, showing the relative decrease in time for the semi-automatic process, compared to the manual one. We observed shorter times for the former process, for all but one of the experiment participants (A3 was slightly faster in the manual task), and an overall time decrease of more than 30%.

Examining the number of laughter events labelled after the two tasks, we noticed, on average, a lower number for the semi-automatic process (78, of which 56 laughs) than for the manual

Table 2: *Per-speaker performance comparison between the semi-automatic and the manual process: relative decrease in time required and the percentage of missed laughter events.*

Annotator	% time decrease	% missed laugh
A1	21.7	5.6
A2	31.1	10.0
A3	-3.3	13.3
A4	33.3	15.4
A5	57.6	9.3
A6	43.8	13.4
A7	7.6	6.4
Overall	30.1	10.2

task (87, with 62 laughs), as expected. Evaluating the difference with a Wilcoxon signed-rank test, a significant result was obtained ($p = .022$). The percentage of laughter events labelled in the manual task and not labelled in the semi-automatic one, by experiment participant, is reported in the last column of Table 2. An average of 10.2% of laughter events were missed in the latter task. In order to better understand the functioning of the automatic tool, we performed an analysis of the types of laughter not found in the semi-automatic process. Of the total of seven events, four were in the form of out-breaths, two laughter interjections (one-syllable laughs) and one a one-syllable speech-laugh. We then examined the annotations corresponding to these events from the manual process, observing a much lower κ value for them (.49 [.30,.69]) than for the whole set (.69), indicating a higher ambiguity of these events.

Finally, we illustrate the inter-rater reliability results for the two tasks in Table 3. It shows the κ and agreement measures when we considered only one laughter class (merging laughs and speech-laughs) or two of them (keeping laughs and speech-laughs separate). We ran Wilcoxon rank sum tests to check the statistical significance of the reported agreement differences between the two annotation processes. None of the tests returned a significant result (two laughter classes: $p = 0.462$, one class: $p = 0.653$). For evaluating the κ measure, we made use of the confidence intervals returned by the tool. According to [23], only an overlap smaller than .5 between confidence intervals would correspond to a p-value indicating statistical significance. The overlap in our case was larger, implying a lack of significant difference also between κ values.

Table 3: κ *measure, its confidence interval (CI), and overall agreement (agr.) for the manual and semi-automatic tasks. One (L+SL) or two (L,SL) laughter event categories were considered (L – laugh, SL – speech-laugh).*

# categories	manual		semi-automatic	
	κ [CI]	agr.	κ [CI]	agr.
1 (L+SL)	.79 [.72,.86]	.897	.77 [.70,.85]	.887
2 (L, SL)	.70 [.64,.76]	.798	.66 [.59,.73]	.771

5. Discussion and conclusions

The current study aimed at establishing two important parameters for the annotation of laughter in speech corpora, independent of other types of annotations. Employing a 15 minutes dyadic conversation in an experiment with seven participants, we observed that the effort needed to complete the task was, on average, five times larger than the duration of the processed file. Of course, these estimates depend on various speaking style characteristics (e.g. the number of pauses produced or the amount of overlapping speech between the dyad partners), as well as on the extent to which laughter is used in the conversation. Here, about two-thirds of the chosen dialogue contained actual speech, with the rest pauses, which makes it one of the dialogues with the least speech among the considered recordings. The proportion of speech in the dialogue is also lower than results published in other studies (e.g. 72%-80% in [24]), suggesting that our average annotation time could be towards the lower bound of the process. With respect to the number of laughter events, the selected dyad produced more laughter occurrences than the average dyad in the Dream Apartment scenario. Their mean of 21.7 laughter events per 10 minutes is higher than the 5.8 rate given by [25], but similar to that observed in other parts of the DUEL corpus [26]. This implies, instead, a higher annotation effort than one could expect for an average dyadic interaction.

We obtained good inter-rater reliability when discriminating laughter events from speech and a slightly worse estimate when a three-way discrimination (laughter events split into laughs and speech-laughs) was considered. Hough and colleagues [9] have previously reported a κ measure of .91 for speech-laughs, on a recording from the Dream Apartment scenario. Our κ value for one laughter category, .79, is lower, but it includes both speech-laugh and laughs, thus also potentially more confounding sources. The level of granularity of our analysis is similar to the bout level of [13], and our overall two laughter classes agreement (.798), falls within the range of "% matched units" they obtained for various conversational corpora (between .63 and .8). This suggests that similar annotation reliability levels can be reached by both experts (as in [13]) and non-experts (as most participants here). We believe that the use of clear annotations schemes for laughter, like the one proposed in [13], would increase the reliability of the process.

The proposed semi-automatic tool produced promising results. Using a signal-based method to pre-select regions where laughter may occur decreased the annotation time by almost a third compared to the manual process. This represents an effort of 3.5 times the length of the considered recording and has a similar rater reliability to the manual task. The observed effect cannot be due to the order in which the annotators performed the tasks as it was balanced (if anything, a manual process advantage should have been seen, as we had four annotators carrying out this task last). We are, thus, confident that these results can be replicated on other datasets. The decrease in work time came at a cost though, a reduction of about a tenth in the number of retrieved laughter events. However, all the missed events were short (only one laughter syllable) and they also featured weak agreements in the manual task. With current automatic detection systems mainly focusing on "prototypical" laughter, their reliability for discovering less common event types remains an open question. Further work would be needed to develop methods able to identify a wider range of laughter types.

6. Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 799022. The authors would like to thank the annotators for their help.

7. References

[1] J. Trouvain and K. P. Truong, "Laughter," in *The Routledge Handbook of Language and Humor*. Routledge, 2017, pp. 340–355.

[2] M. Mancini, R. Niewiadomski, S. Hashimoto, M. E. Foster, S. Scherer, and G. Volpe, "Guest editorial: Towards machines able to deal with laughter," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 492–494, 2017.

[3] H. Rothgänger, G. Hauser, A. C. Cappellini, and A. Guidotti, "Analysis of laughter and speech sounds in Italian and German students," *Naturwissenschaften*, vol. 85, no. 8, pp. 394–402, 1998.

[4] Y. Tian, C. Mazzocconi, and J. Ginzburg, "When do we laugh?" in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 360–369.

[5] B. Ludusan and P. Wagner, "Laughter dynamics in dyadic conversations," in *Proceedings of INTERSPEECH*, 2019, pp. 524–528.

[6] B. Ludusan, M. Wesemann, and P. Wagner, "A distributional analysis of laughter across turns and utterances," in *Proceedings of the Laughter and Other Non-Verbal Vocalisations Workshop*, 2020.

[7] J. Hough, Y. Tian, L. de Ruiter, S. Betz, S. Kousidis, D. Schlangen, and J. Ginzburg, "DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter," in *Proceedings of the 10th Language Resources and Evaluation Conference*, 2016, pp. 1784–1788.

[8] K. P. Truong and J. Trouvain, "Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis," *Proceedings of the 4th International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals*, pp. 20–24, 2012.

[9] J. Hough, L. de Ruiter, S. Betz, and D. Schlangen, "Disfluency and laughter annotation in a light-weight dialogue mark-up protocol," in *Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech*, 2015, pp. 49–52.

[10] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image and Vision Computing*, vol. 31, no. 2, pp. 186–202, 2013.

[11] K. Laskowski and S. Burger, "On the correlation between perceptual and contextual aspects of laughter in meetings," *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, pp. 55–60, 2007.

[12] C. Ishi, H. Hatano, and N. Hagita, "Analysis of laughter events in real science classes by using multiple environment sensor data," in *Proceedings of INTERSPEECH*, 2014, pp. 1043–1047.

[13] K. P. Truong, J. Trouvain, and M.-P. Jansen, "Towards an annotation scheme for complex laughter in speech corpora," in *Proceedings of INTERSPEECH*, 2019, pp. 529–533.

[14] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*, 2013, pp. 148–152.

[15] G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André, "The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 166–172.

[16] S. Kipper and D. Todt, "The role of rhythm and pitch in the evaluation of human laughter," *Journal of Nonverbal Behavior*, vol. 27, no. 4, pp. 255–272, 2003.

[17] L. Varnet, M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1976–1989, 2017.

[18] B. Ludusan and P. Wagner, "Speech, laughter and everything in between: A modulation spectrum-based analysis," in *Proceedings of Speech Prosody*, 2020, pp. 995–999.

[19] L. Varnet, "Matlab toolbox for the computation of amplitude- and frequency- modulation spectra," https://github.com/LeoVarnet/AM_FM_Spectra, 2018.

[20] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2002.

[21] J. J. Randolph, "Free-marginal multirater kappa (multirater κ [free]): An alternative to Fleiss' fixed-marginal multirater kappa," in *Joensuu Learning and Instruction Symposium*, 2005.

[22] ——, "Online kappa calculator," [Computer software]. Retrieved from http://justusrandolph.net/kappa/, 2008.

[23] G. Cumming and S. Finch, "Inference by eye: Confidence intervals and how to read pictures of data." *American Psychologist*, vol. 60, no. 2, pp. 170–180, 2005.

[24] K. Lundholm Fors, "Production and perception of pauses in speech," Ph.D. dissertation, University of Gothenburg, 2015.

[25] J. Vettin and D. Todt, "Laughter in conversation: Features of occurrence and acoustic structure," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 93–115, 2004.

[26] C. Mazzocconi, Y. Tian, and J. Ginzburg, "Multi-layered analysis of laughter," in *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue*, 2016, pp. 97–107.