10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Speech, laughter and everything in between: A modulation spectrum-based analysis

Bogdan Ludusan, Petra Wagner

Phonetics Workgroup, Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
CITEC, Bielefeld University, Germany
{bogdan.ludusan, petra.wagner}@uni-bielefeld.de

Abstract

Laughter and speech-laughs are pervasive phenomena found in conversational speech. Nevertheless, few previous studies have compared their acoustic realization to speech. We investigated in this work the suprasegmental characteristics of these two phenomena in relation to speech, by means of a modulation spectrum analysis. Two types of modulation spectra, one encoding the variation of the envelope of the signal and the other one its temporal fine structure, were considered. Using a corpus of spontaneous dyadic interactions, we computed the modulation index spectrum and the f0 spectrum of the three classes of vocalizations considered and we fitted separate generalized additive mixed models for them. The results obtained for the former modulation showed a clear separation between speech, on the one hand, and laughter and speech-laugh, on the other hand, while the f0 spectrum was able to discriminate between all three classes. We conclude with a discussion of the importance of these findings and their implication for laughter detection.

Index Terms: laughter, speech-laugh, speech, modulation spectrum, paralinguistics

1. Introduction

Among the non-verbal vocalizations employed in spontaneous human-human interactions one can often encounter laughter and speech-laughs [1]. Laughter is a phenomenon generally produced as one or several rhythmic "bouts", representing sequences of laughter syllables (calls) uttered during an exhalation phase, separated by an inhalation part [2]. A typical laughter contains consonant-vowel-like calls, with the consonantal part being an aspirated consonant. Due to these characteristics, earlier laughter studies have considered laughter structure as being highly stereotypical (e.g. [3]). Nevertheless, more recent acoustic studies have shown that laughter actually displays a high degree of intra- and inter-individual variability (e.g. [4, 5]). Despite this variability, laughter has been shown to be one of the most recognizable emotional vocalizations, even across cultures [6].

What makes laughter so distinguishable? It might be that its acoustic characteristics set it apart from other vocalizations and even from speech. In a couple of studies, Kipper and Todt [7, 8] evaluated the perceptual quality of laughter when its rhythm and fundamental frequency (f0) patterns were varied. They found that the evaluation of laughter depends both on the rhythmic structure of the bouts and the dynamic changes of the acoustic parameters within laughter calls. They hypothesized that laughter evaluation involves two steps: the first one uses basic properties of laughter, such as its rhythmic structure, to make a decision whether a succession of vocal elements might represent laughter. If a positive decision is taken, the second step will determine the quality of the laughter by means of evaluating parameter variations within successive laughter calls (e.g. f0 variation). Further evidence regarding the distinct rhythm exhibited by laughter has been found in investigations looking at air flow, pressure and muscle activity in the case of laughter (see [9] for a review), while the specific values of f0 for laughter have been examined in numerous acoustic studies (e.g. [10, 3, 4, 11, 5]), showing that laughter displays higher f0 mean values and ranges than speech.

Speech-laughs represent concurrent productions of speech and laughter, in which neither of the two components is dominant. Although they occur often in conversational speech [11, 12], they have been less studied than laughter, with only a small number of investigations having directly compared the acoustic characteristics of speech-laughs, laughter and speech (e.g. [11, 13, 14]). According to [11], speech-laughs exhibit speech-like fundamental frequency and laughter-like rhythm and amplitudes, while [13] showed that the spectral envelope of speech-laughs is a combination of speech and laughter envelopes. Also [14], looking at several measures characterizing the following excitation source components: f0, rate of glottal closure, and opening of the glottal folds, found values for speech-laughs falling between those for laughter and speech.

In this study, we will investigate the two prosody components previously introduced, rhythm and f0 variation, by employing information extracted from two pertinent representations. The speech signal may be viewed as a modulated carrier signal (e.g. [15]), composed of an amplitude modulation (AM) and a frequency modulation (FM) component, respectively. The former encodes the variations of the temporal envelope of the signal, while the latter relates to the temporal fine structure of the signal. It has been previously shown that the amplitude envelope of the signal and a closely linked description, the AM spectrum, are able to discriminate between language classes [16, 17]. Furthermore, information about the f0 variation can be extracted from the f0 spectrum, itself a sub-part of the FM component [17].

We examine here the role the AM and f0 spectra play in discriminating between laughter, speech-laughs and speech. We expect the AM spectrum to be able to differentiate between laughter and speech, based on the rhythmic characteristic of the former. Also, the f0 spectrum may be able to discriminate between laughter and speech, seeing how the two classes differ in average f0 and f0 range, but it may differentiate the two classes less than the AM spectrum. No clear hypotheses can be formulated for the discrimination between speech-laughs, on the one hand, and laughter and speech, on the other hand, except that the former should have characteristics somewhere in between those of the latter two.

10.21437/SpeechProsody.2020-203

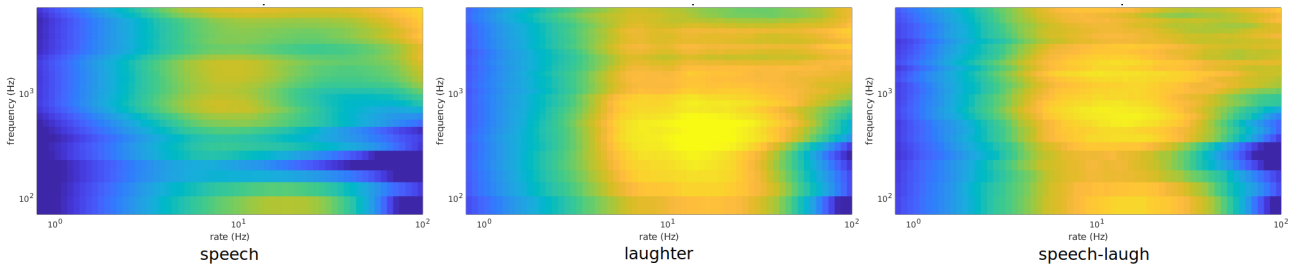

Figure 1: *Average modulation index spectrum obtained for the three considered classes: speech, laughter and speech-laugh. The vertical axis represents the audio frequency, while on the horizontal axis we have the modulation rate.*

2. Materials

The DUEL corpus [18] was employed for the experiments conducted in this study. It contains spontaneous interactions between dyads asked to discuss one of three possible scenarios: "Dream Apartment", "Film Script" and "Border Control", with recordings being made in German, French and Mandarin Chinese. The materials contain a large number of phenomena occurring in conversational speech, such as disfluencies, hesitations or laughter. The corpus was fully transcribed and segmented at the utterance level, having also manual annotations for the previously mentioned spoken language phenomena.

We chose here a subset of the German data, consisting of eight dyads discussing the "Film Script" scenario. In this case, the speakers were asked to come up with the script for a film consisting of an embarrassing moment, with personal experience being a possible inspiration source. All but one of the considered dyads were composed of friends/colleagues, with the dyads containing a total of 11 females and 5 males. The selected subset consisted of almost two hours of recordings.

Each recording was then divided into the three classes of vocalizations we investigated in this study: speech, laughter and speech-laughs, based on the manual annotations. We discarded utterances containing occurrences of more than one element class, as they might confound our analysis. Among the three classes, we found that the duration of the laughter class instances was significantly shorter (922 ms) than that of instances occurring in the speech and speech-laugh classes (1510 ms and 1585 ms, respectively). Since it has been observed that the length of the analysed stimulus has an impact on the shape of the modulation spectrum [17], we decided to mitigate this risk, by concatenating shorter laughter instances belonging to the same speaker. The procedure checked whether shorter (non necessarily consecutive) laughs can be concatenated into a laughter sequence that did not exceed the average length of the speech class occurrences. This process resulted in laughter stimuli, having a similar duration, on average, to that of the other two classes (1520 ms). In total, 2964 instances of speech, 390 occurrences of laughter and 171 speech-laughs have been analysed. A final pre-processing step saw the root-mean-square (RMS) normalization of the investigated stimuli, performed on a per-speaker basis.

3. Methods

The elements of each of the three classes of vocalizations had their spectra computed by means of a Matlab toolbox [19], implementing the modulation spectra investigated in [17]. Of the four types of spectra computed by the toolbox, we employ one type of AM modulation (called modulation index spectrum - $ModSp$) and the f0 spectrum ($f0Sp$). The modulation index spectrum was chosen over the standard AM spectrum as it does not represent an absolute value, but a ratio between the intensity of the signal and the noise, being more closely linked to human perception performance. Furthermore, we focused our investigation on the f0 spectrum, since one might expect a difference between the three classes in this respect, seeing how f0 differences have been observed between speech and laughter (e.g. [10, 11]).

The two spectra were computed as follows: In order to obtain the $ModSp$ spectrum, the signal was first band-pass filtered using a gammatone filterbank with 30 channels. Each channel was 1 ERB-wide and their center frequencies were equally spaced on the ERB scale. The envelopes of the resulting signals were extracted from the response of each filter channel and each envelope was then filtered using a bank of $1/3$ octave wide Butterworth bandpass filters overlapping at -3 dB. The RMS amplitude of each filter output was multiplied by $\sqrt{2}$ and the corresponding modulation index was obtained by dividing this output by the mean amplitude of the output of the gammatone filter. The average across the 30 channels was taken, in order to obtain a single modulation index spectrum. For the calculation of the $f0Sp$ spectrum, the f0 was first extracted, by considering only voiced segments longer than 20 ms and having an f0 range between 50 and 550 Hz. Once the f0 from all the stimuli was extracted, the root of the Lomb periodogram (a generalization of the Fourier spectrum for partially undefined functions) was employed to compute the f0 spectrum. In this study, we limited our analysis to the [0, 30] Hz modulation rate band, as the modulation spectra of speech contain most of their energy in this range [17].

To investigate potentially vocalization type-induced non-linear effects in the the modulation spectra as a function of the modulation rate ($rate$), we entered the time-normalized spectra $ModSp$ and $f0Sp$ as dependent variables into generalized additive mixed models (GAMMs), using the R-package *mgcv* [20], and following the procedures suggested in [21]. We used thin plate regression splines as smooths to model the non-linear variation present in the data, and checked for potential over-smoothing. We entered vocalization type as a fixed factor with 3 levels into the models, thereby determining whether vocalization type significantly influences the non-linear shape of the modulation spectra:

1. SP = speech
2. SL = speech-laugh
3. LG = laughter

As we expect individual speakers to have an impact on the shape of the spectra of the three vocalization types, we added

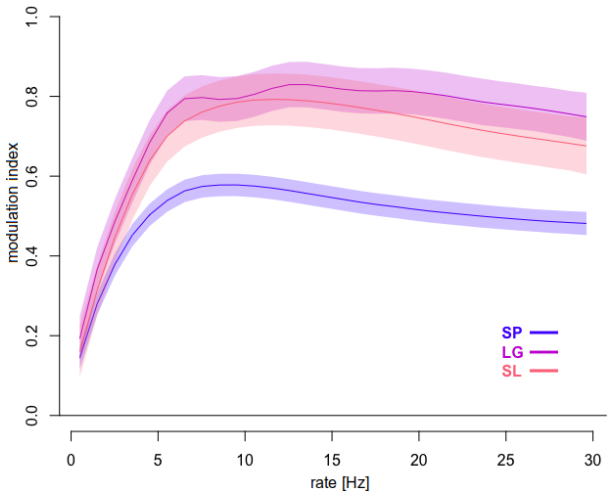

Figure 2: *Fitted values for the modulation index spectrum, for the three classes of vocalizations: speech (SP), laughter (LG) and speech-laugh (SL).*

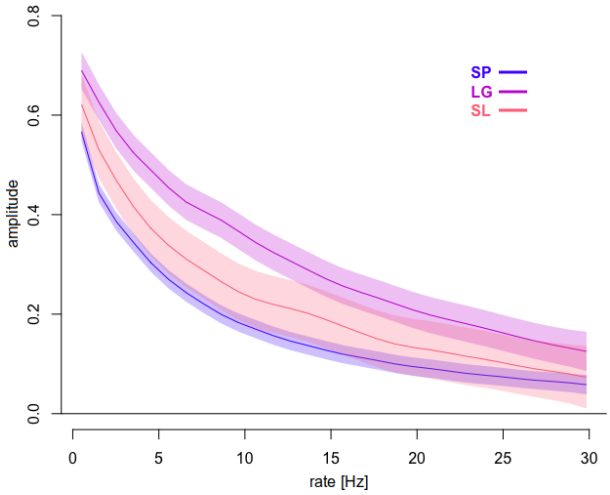

Figure 3: *Fitted values for the f0 spectrum, for the three classes of vocalizations: speech (SP), laughter (LG) and speech-laugh (SL).*

a non-linear random factor smooth in our models to control for this. More specifically, we entered three vocalization-type specific smooths per speaker as a function of rate. We tested the resulting model against a base model not containing the fixed factor using the function *compareML* of the R-package *itsadug* [22]. To determine the regions of significant differences between the non-linear smooths for the various factor levels, we estimated pairwise differences between the confidence bands for non-linear smooths. Then, we considered as significantly different those regions across the spectra where the difference between confidence bands differs from zero.

4. Results

In Figure 1 we illustrate the average modulation index spectrum obtained for each of the three classes, based on all the data we included in the study. One can observe important differences in the modulation index amplitude distribution across frequencies, between the speech and laughter classes. While the former presents medium-to-high modulation indices in the frequency bands around 100 Hz and 1000 Hz, for modulation rates around 10 Hz, the latter has very high modulation indices in the entire frequency range 0-1000 Hz and for a wider range of modulation rates than in the case of speech signals. The spectrum of the speech-laugh class seems to resemble that of the laughter class, while also exhibiting characteristics of the spectrum of the speech class (lower modulation indices in the 200-800 Hz band and a peak value around 1000 Hz).

For both the modulation index spectrum and the f0 spectrum, a model comparison revealed highly significant differences to a base model that does not contain the fixed factor *vocalization type*.

The model predicting ModSp (adjusted $r^2 = 0.41$), shows a highly significant impact of vocalization type on the intercept for laughter in comparison to speech, but not to speech-laugh (upper part in Table 2), and a highly significant impact of smooth terms for fixed (vocalization type over rate, middle part in Table 1) and random factors (vocalization type specific speaker-wise variation over rate, lower part in Table 1). Note that the edf-value (effective degrees of freedom) indicates the amount of non-linearity in the smooth, while the ref.df value gives the number of reference degrees of freedom used for hypothesis testing in relation to the associated F-value. An illustration of the model smooths across vocalization types can be found in Figure 2. A direct comparison of the smooths for the different vocalization types revealed significant differences be-

Table 1: *Model overview for ModSp*

intercepts	*estimates*	*SE*	*t*	*p*
LG	0.572	0.024	24.1	***
SL	-0.04	0.038	-1.17	n.s.
SP	-0.16	0.027	-6.11	***
smooth terms	*edf*	*ref.df*	*F*	*p*
fixed				
s(rate):LG	14.24	16.53	60.98	***
s(rate):SL	10.64	12.95	92.42	***
s(rate):SP	16.46	18.15	170.56	***
random				
s(rate,spkr):LG	98.90	143	34.60	***
s(rate,spkr):SL	52.96	107	21.09	***
s(rate,spkr):SP	106.29	143.0	68.67	***

Table 2: *Model overview for f0Sp.*

intercepts	*estimates*	*SE*	*t*	*p*
LG	0.432	0.016	27.9	***
SL	-0.10	0.027	-3.57	***
SP	-0.16	0.017	-9.34	***
smooth terms	*edf*	*ref.df*	*F*	*p*
fixed				
s(rate):LG	10.86	13.10	50.32	***
s(rate):SL	11.31	13.56	15.10	***
s(rate):SP	18.39	18.89	148.07	***
random				
s(rate,spkr):LG	99.24	143	27.86	***
s(rate,spkr):SL	81.59	107	17.09	***
s(rate,spkr):SP	120.12	143.0	51.27	***

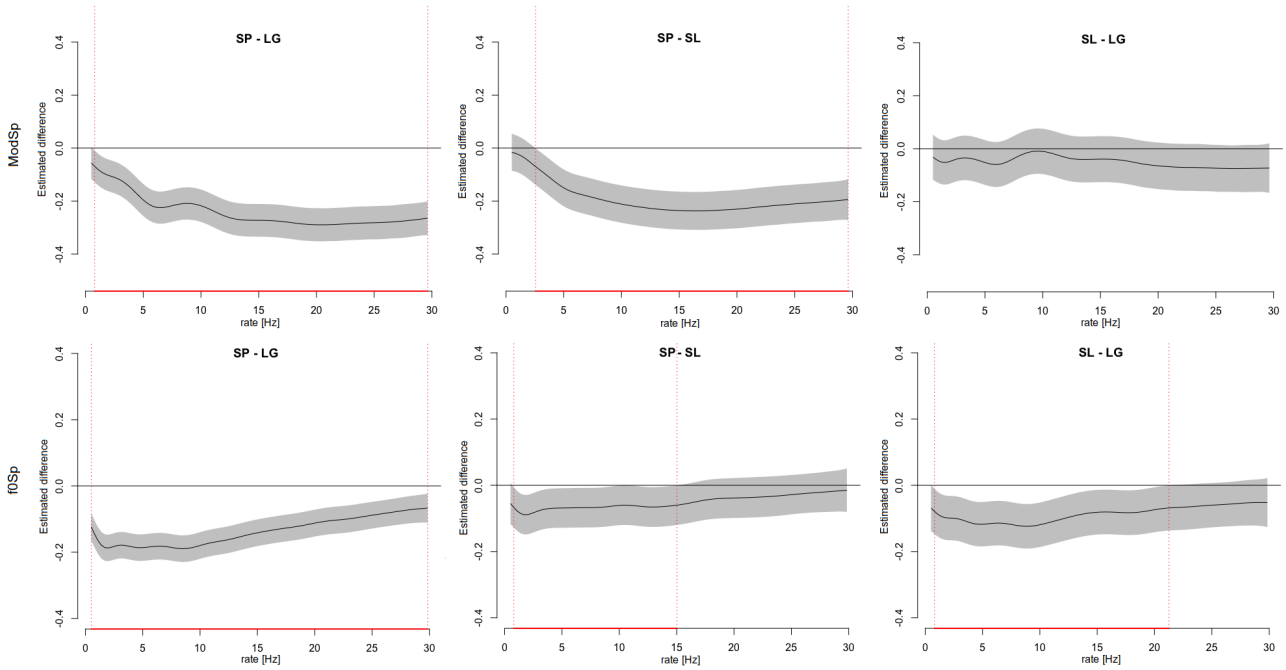

Figure 4: *Differences between the fitted class models for the modulation index spectrum (top panels) and f0 spectrum (bottom panels), for the three classes: speech (SP), laughter (LG) and speech-laugh (SL). The red interval on the horizontal axis represents the modulation rate range for which the two models differ significantly.*

tween speech and both laughter and speech-laugh across almost the entire range of analysed modulation rates, but no significant differences between laughter and speech-laugh. (Figure 4, upper panel).

The model predicting f0Sp (adjusted $r^2 = 0.49$), shows a highly significant impact of vocalization type on the intercepts (upper part in Table 2, and a highly significant impact of smooth terms for fixed (vocalization type over rate, middle part in Table 2) and random factors (vocalization type specific speaker-wise variation over rate, lower part in Table 2). Figure 3 illustrates the model smooths across the different vocalization types. A comparison of the smooths, across vocalization types, revealed significant differences between all vocalization types, for a wide range of modulation rates (Figure 4, lower panel).

5. Discussion and conclusions

We have seen in the previous section that the obtained results support the hypotheses put forward in the introduction. First, the AM representation employed in this study clearly differentiates between laughter and speech. Laughter exhibits different peak rates in its $ModSp$ spectrum compared to speech (several peaks rates, all of them significantly higher in amplitude than the sole peak rate obtained for speech). It did not discriminate, however, between laughter and speech-laugh, in line to the findings of [11] that the two phenomena share a similar rhythm. Second, the f0 spectrum information was successful in separating the laughter class from the speech class. Moreover, and somewhat surprisingly, f0 variation was able to discern between the speech-laugh class and the other two vocalization classes considered here.

Our findings back also the hypothesis proposed in [7], regarding the recognition of laughter. Rhythm information, extracted from the the modulation index spectrum, in conjunction with information on f0 variation, obtained from the f0 spectrum, may be used towards this goal. Furthermore, our results point towards a possible use of the same mechanism also for the discrimination between speech-laughs and speech. The modulation spectrum can be used to discriminate laughter and speech-laugh from speech, while information extracted from the f0 spectrum can be employed to differentiate between the former two classes.

We have seen in this study how suprasegmental information (related to rhythm and f0) may be used to discriminate between laughter, speech-laugh and speech. Taking into account that laughter exhibits high inter- and intra-speaker variation, the studied components seem to display an important characteristic, that of being robust to such variations. These findings encourage us to consider this type of information in automatic laughter detection systems. Although outside the scope of this study, we can envisage the use of such information, not only in supervised systems (e.g. [23]), but also in systems not employing any type of learning (e.g. in a system implementing the laughter evaluation procedure proposed in [7]). In the future, we would like to explore also other prosodic features, such as intensity and its variation, that might be useful for the discrimination of these vocalization classes.

6. Acknowledgements

Bogdan Ludusan's work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 799022.

7. References

[1] J. Trouvain and K. P. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proceedings of the*

LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals, 2012, pp. 36–39.

[2] J. Trouvain, "Segmenting phonetic units in laughter," in *Proceedings of the 15th International Congress of Phonetic Sciences*, 2003, pp. 2793–2796.

[3] R. R. Provine and Y. L. Yong, "Laughter: A stereotyped human vocalization," *Ethology*, vol. 89, no. 2, pp. 115–124, 1991.

[4] 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.

[5] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[6] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2408–2412, 2010.

[7] D. Todt and S. Kipper, "Variation of sound parameters affects the evaluation of human laughter," *Behaviour*, vol. 138, no. 9, pp. 1161–1178, 2001.

[8] 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.

[9] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotions, qualia, and consciousness*. World Scientific, 2001, pp. 426–443.

[10] D. E. Mowrer, L. L. LaPointe, and J. Case, "Analysis of five acoustic correlates of laughter," *Journal of Nonverbal behavior*, vol. 11, no. 3, pp. 191–199, 1987.

[11] E. E. Nwokah, H.-C. Hsu, P. Davies, and A. Fogel, "The integration of laughter and speech in vocal communication: A dynamic systems perspective," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 4, pp. 880–894, 1999.

[12] B. Ludusan and P. Wagner, "Laughter Dynamics in Dyadic Conversations," in *Proceedings of INTERSPEECH*, 2019, pp. 524–528. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1733

[13] C. Menezes and Y. Igarashi, "The speech laugh spectrum," *Proceedings of the 7th International Seminar on Speech Production, Brazil*, pp. 157–164, 2006.

[14] S. H. Dumpala, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Use of vowels in discriminating speech-laugh from laughter and neutral speech." in *Proceedings of INTERSPEECH*, 2016, pp. 1437–1441.

[15] R. Plomp, "The role of modulation in hearing," in *Hearing—Physiological bases and psychophysics*. Springer, 1983, pp. 270–276.

[16] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, 2013.

[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] 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.

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

[20] S. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[21] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, vol. 70, pp. 86 – 116, 2018.

[22] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using GAMMs," 2017, r package version 2.3.

[23] K. P. Truong and D. A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.