Laughter and Other Non-Verbal Vocalisations Workshop 2020
5 October 2020, Bielefeld, Germany

Comparing Annotations of Non-verbal Vocalisations in Speech Corpora

Jürgen Trouvain
Saarland University
Language Science and Technology
Saarbrücken, Germany
trouvain@lst.uni-saarland.de

Raphael Werner
Saarland University
Language Science and Technology
Saarbrücken, Germany
rwerner@lst.uni-saarland.de

Abstract

In this study eleven corpora of spontaneous and scripted speech (in English and in German) are analysed regarding their annotation inventories of selected highly frequent non-verbal vocalisations (NVVs). It appears that only one corpus considers all NVVs and that laughter is the only NVV annotated in all corpora. The findings lead to a discussion of possible reasons for this situation. In conclusion it is argued that a wider distribution and more consistency is needed with respect to the annotation of NVVs.

1 Introduction

The investigation of non-verbal vocalisations (NVVs) such as laughter and syllable-like forms like fillers ('filled pauses') is often based on data elicited in experiments. Another rich source of data to investigate NVVs are corpora of various other speech modes. Although these databases are not recorded with NVVs as explicit research objects they usually take into account that not only words, or verbal vocalisations, can have importance in the speech signal. Breath noises are a typical example of an NVV that occurs in all kinds of speech styles including types of scripted and otherwise prepared speech. However, it is unclear whether and how NVVs like breath noises are annotated in speech corpora.

The aim of this exploratory study is to review and compare selected types of NVVs in annotations in speech corpora that consist of spontaneous as well as scripted speech. Although we can expect that spontaneous dialogues contain more NVVs than scripted speech, which is often represented by isolated sentences, all kinds of scripted speech are expected to contain breath noises.

2 Corpora

A total of eleven corpora are investigated here. Of those, six corpora with conversational English were already studied in a previous paper (Trouvain and Truong, 2012):

- ICSI meeting corpus (Janin et al., 2003),
- AMI (Carletta, 2007),
- Switchboard (Godfrey and Holliman, 1997),
- Diapix Lucid corpus (Baker and Hazan, 2011),
- HCRC Map Task corpus (Anderson et al., 1991),
- Buckeye corpus (Pitt et al., 2007).

This set was complemented by five German corpora, the first two of which consist of dialogues, the other three contain scripted speech:

- GECO (Schweitzer and Lewandowski, 2013),
- Lindenstrasse corpus (IPDS, 2006),
- Kiel Corpus of Read Speech (IPDS),
- IFCASL corpus (Trouvain et al., 2016),
- DIRNDL (Björkelund et al., 2014).

3 Annotation categories

In this study only highly frequent types of NVVs have been considered. It should be mentioned that most of the selected NVVs are produced as phonetic particles in speech pauses. Often pauses are divided into 'filled pauses' and 'silent' pauses, both of which are established technical terms in phonetics and speech fluency research. However, 'filled pauses' usually refer to filler syllables (and not

to pauses) and the majority of 'silent pauses' do contain breath noises, not only portions of silence (Trouvain and Belz, 2019). In this light it is worth mentioning that only one corpus, the Buckeye corpus (Pitt et al., 2007), makes a distinction between a pause on the one hand and silence on the other.

3.1 Laughs and speech-laughs

Laughs and speech-laughs are not necessarily NVV categories that can be expected in scripted speech. For this reason they are not used in two of the three inspected corpora of scripted speech (see Table 1 for an overview of the annotation categories of the selected NVVs in the inspected corpora).

Laughter as a typical phenomenon of spontaneous speech is annotated in all corpora of spontaneous speech. Laughter in speech communication usually also affects spoken sections, i.e., speech-laughs, in which speech and laughing happen at the same time. Speech-laughs represent an extra category of laughter and require an annotation that mark the speech-laughed sections. However, there are two corpora in which speech-laughs are not annotated as such an extra category.

3.2 Breath noise

The category 'breath noise' is the most frequent NVV in speech. This statement is valid for spontaneous conversations (Trouvain and Truong, 2012) and can also be assumed for scripted speech. Three spontaneous corpora and one scripted corpus are not annotated for 'breath noises'.

3.3 Coughing and throat clearing

Although 'cough' is an infrequent NVV category in comparison to 'breath noise', a fair amount can be found in spontaneous speech. Three spontaneous corpora do not include annotations for 'cough'.

Coughing can certainly also occur in scripted speech. However, here researchers and corpus providers are usually interested in having 'clean' data, i.e., a cough in a recording session leads either to a repetition of the sentence or text by the recorded speaker or is excluded from the published data. It might be for these reasons that none of the scripted speech corpora have coughing as an NVV category of its own.

Similar to coughing is throat clearing. Only four corpora make use of this NVV category (one of which is a scripted speech corpus).

3.4 Clicks and lip smacking

It is a rather recent discovery that languages that do not have clicks as phonemes such as English and German show a fairly frequent usage of (probably unconsciously produced) tongue clicks (Trouvain, 2014). Interestingly, in some corpora there is a considerable amount of 'lip smacks', a sound class in which also (tongue) clicks could be easily subsumed. Further research is needed on the production mechanisms of these sounds and whether (some) 'lip smacks' are in reality tongue clicks, and also vice versa, whether some tongue clicks may be labially articulated. This common NVV category is neglected in six of the eleven corpora.

4 Discussion and summary

In summary, NVVs seem to have a 'Cinderella status' in the annotation of speech corpora, since many of them are neglected in various corpora. Only one of the eleven corpora shows all selected NVVs. Interestingly, for the spontaneous corpora, the only NVV used by all is laughter.

Breath noises, which are definitively more frequent than laughs, are not annotated in all corpora. One reason might be that most breath noises show a much softer intensity and thus might go unnoticed by those researchers who set up the annotation categories. As previously mentioned most so-called 'silent' pauses contain breath noises (Trouvain and Belz, 2019), i.e., making 'silent' pauses 'non-silent' - a fact that supports the idea of a lack of attention for this NVV.

Most of the selected NVVs are represented by one category only without a further subcategorisation. The exception is laugh and speech-laugh. Laughs in general can have an immense variability regarding their form and complexity. In theory it could be beneficial to annotate laughs with further sub-units such as episodes, calls, syllable-like units and smaller segments. However, the development of an appropriate annotation scheme is challenging (Truong et al., 2019).

Breath noises could be sub-divided into inhalation and exhalation noises, and potentially further into oral and nasal breath noises. However, such a detailed distinction is not always reliably possible when annotating corpora (Trouvain and Belz, 2019).

The existence of a given annotation category does not indicate that this category was used in all cases or that the temporal alignment follows sim-

Table 1: Overview of the selected NVVs and whether they were annotated in the inspected corpora.

	Laugh	**Speech-laugh**	**Breath**	**Cough**	**Throat clearing**	**Clicks & lip smacks**
spontaneous						
ICSI	y	y	y	y	y	n
AMI	y	n	y	y	n	y
Switchboard	y	y	n	n	n	n
Diapix	y	y	y	n	n	y
MapTask	y	n	y	y	n	y
Buckeye	y	y	n	n	n	n
GECO	y	y	n	y	y	n
Lindenstr	y	y	y	y	y	y
scripted						
KielCorpusRead	y	y	y	n	y	y
IFCASL	n	n	y	n	n	n
DIRNDL	n	n	n	n	n	n

ilar standards. In a recent paper by Zayats et al. (Zayats et al., 2019) transcription errors of human labellers were analysed for the Switchboard corpus (Godfrey and Holliman, 1997). Unexpectedly, they found relatively high error rates for filler syllables, interjections and short feedback expressions. The authors speculate that the high error rates they found may be due to a relatively "low information load and/or lack of conscious awareness of spontaneous speech phenomena. Lack of awareness would explain the need for annotator training. Further study is needed." The evidence of the small study presented here fully supports this opinion.

Thus, it would be a mid-term goal for the community to develop a common set of annotation guidelines, at least to raise the awareness for the mentioned inconsistencies. It could be a start to investigate how consistent the available annotations of the various NVVs in question actually are - in more than just the corpora presented here and across many languages and speech modes. This would also help discover (and ideally understand) potentially subtle phenomena. Apart from finding commonalities among corpora and suggesting labels for a standardisation of NVVs we need more research on some theoretical concepts. This mainly concerns those phonetic particles that occur in speech pauses. For instance, breath noises can show a huge variability but also a multitude of functions, from marking a syntactic-prosodic break up to signalling arousal (Trouvain et al., 2020).

A helpful technique in finding and annotating NVVs are automatic or semi-automatic detection procedures. There are useful approaches for instance for breath noise detection (see e.g. (Braunschweiler and Chen, 2013; Fukuda et al., 2018; Székely et al., 2019)) but there is also a need for a systematic approach to the automatic detection of *all* important and frequent NVVs.

5 Acknowledgments

We would like to thank Bernd Möbius and two anonymous reviewers for their helpful recommendations.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Rachel Baker and Valerie Hazan. 2011. DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3):761–770.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 3222–3228.

Norbert Braunschweiler and Langzhou Chen. 2013. Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *8th ISCA Workshop on Speech Synthesis*, July, pages 1–6.

Jean Carletta. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190.

Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura. 2018. Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. *Speech Communication*, 98:95–103.

John J. Godfrey and Edward Holliman. 1997. Switchboard-1 Release 2.

IPDS. The Kiel Corpus of Read Speech (Volume 1, DVD 1). Technical report, Institut für Phonetik und Digitale Sprachsignalverarbeitung, University of Kiel.

IPDS. 2006. Video Task Scenario: Lindenstraße – The Kiel Corpus of Spontaneous Speech. DVD 4, Institut für Phonetik und Digitale Sprachsignalverarbeitung Universität Kiel.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. THE ICSI MEETING CORPUS. *Proceedings of ICASSP*, pages 364–367.

M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymon, E. Hume, and E. Foster-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release).

Antje Schweitzer and Natalie Lewandowski. 2013. Convergence of articulation rate in spontaneous speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August):525–529.

Éva Székely, Gustav Eje Henter, and Joakim Gustafson. 2019. Casting to Corpus: Segmenting and Selecting Spontaneous Dialogue for TTS with a CNN-LSTM Speaker-dependent Breath Detector. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6925–6929.

Jürgen Trouvain. 2014. Laughing, Breathing, Clicking - The Prosody of Nonverbal Vocalisations. *Proceedings of the International Conference on Speech Prosody*, pages 598–602.

Jürgen Trouvain and Malte Belz. 2019. Zur Annotation nicht-verbaler Vokalisierungen in Korpora gesprochener Sprache. In *Proceedings 30th Conference Elektronische Sprachsignalverarbeitung (ESSV '19)*, pages 280–287, Dresden.

Jürgen Trouvain, Anne Bonneau, Vincent Colotte, Camille Fauth, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella, Bernd Möbius, and Frank Zimmerer. 2016. The IFCASL corpus of French and German non-native and native read speech. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 1333–1338.

Jürgen Trouvain, Bernd Möbius, and Raphael Werner. 2020. On Acoustic Features of Inhalation Noises in Read and Spontaneous Speech. In *Proceedings 10th International Conference on Speech prosody*, Tokyo.

Jürgen Trouvain and Khiet P Truong. 2012. Comparing non-verbal vocalisations in conversational speech corpora. In *Proceedings of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 36–39, Istanbul.

Khiet P Truong, Jürgen Trouvain, and Michel-Pierre Jansen. 2019. Towards an annotation scheme for complex laughter in speech corpora. In *Proceedings Interspeech 2019*, pages 529–533, Graz.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-Septe:3088–3092.