Detection of nonverbal vocalizations using Gaussian Mixture Models: looking for fillers and laughter in conversational speech

Teun F. Krikke, Khiet P. Truong

Human Media Interaction, University of Twente
Enschede, The Netherlands
{t.f.krikke, k.p.truong}@utwente.nl

10.21437/Interspeech.2013-59

Abstract

In this paper, we analyze acoustic profiles of fillers (i.e. filled pauses, FPs) and laughter with the aim to automatically localize these nonverbal vocalizations in a stream of audio. Among other features, we use voice quality features to capture the distinctive production modes of laughter and spectral similarity measures to capture the stability of the oral tract that is characteristic for FPs. Classification experiments with Gaussian Mixture Models and various sets of features are performed. We find that Mel-Frequency Cepstrum Coefficients are performing relatively well in comparison to other features for both FPs and laughter. In order to address the large variation in the frame-wise decision scores (e.g., log-likelihood ratios) observed in sequences of frames we apply a median filter to these scores, which yields large performance improvements. Our analyses and results are presented within the framework of this year's Interspeech Computational Paralinguistics sub-Challenge on Social Signals.

Index Terms: nonverbal vocalizations, laughter, filled pauses, detection

1. Introduction

Human speech contains a wealth of information about the speaker's emotional, interpersonal, and cognitive states (among others) that are continuously being evaluated during social conversational interaction. This type of information particularly lies in the channel that goes beyond the content of words, i.e., the paralinguistic channel. Paralinguistic information in speech is mostly concerned with feature representations of F0, intensity, speech rate, and voice quality measures. Non-verbal vocalizations, word-like sounds that do not have a clear lexical content, are also part of this paralinguistic space. Examples of relatively distinct non-verbal vocalizations that are especially salient in spontaneous conversational speech are fillers and laughter. Our interest lies in analysing these vocalizations in conversation in order to advance technology that aims to recognize and understand human social and affective behavior in interaction. In this paper, we will analyse fillers and laughters, and develop detectors for these vocalizations.

In recent years, the detection of these types of nonverbal vocalizations have become increasingly important in the community of social signal processing and affective computing. Fillers and laughter can signal important speaker state information in social discourse. A common type of fillers, filled pauses such as 'ehm, uh' are often associated with the speaker's cognitive state and occur often when the speaker is experiencing some sort of increased cognitive load (e.g., [1, 2]). Fillers are also used as mechanisms to maintain the floor [2, 3]. Spoken dialog systems could hence benefit from the detection of fillers. Laughter is often associated with positive attitudes and affilitation. There are many forms of laughter (i.e., chuckle, song-like, etc.) as well as possible functions (i.e., evil laughter, shy laughter, etc.) of laughter. In addition to these speaker-state related descriptions, laughter may also play a more discourse-oriented role in conversation, indicating a topic-change or a way to mitigate the following message. In order to interpret what kind of information these fillers and laughter events yield on a higher level, detection of these vocalizations must first take place.

With the availability of a large corpus of annotated fillers and laughter events, this year's Interspeech Computational Paralinguistics Challenge [15] offers an opportunity for researchers to analyse these vocalizations on a large scale and to compare results in a more controlled way. We take this opportunity and develop methods for the automatic detection of fillers and laughter (excluding speech-laughs) in conversational speech. In contrast to a brute-force data-driven approach, we opt for a more selective aproach where we work with a (relatively) small set of features that is selected based on our insights and previous literature. We introduce the use of voice quality features for laughter detection (which have not often been used for laughter detection) to capture the differences in production modes and the use of spectral similarity features for filler detection. Based on observations in the literature, we find that Gaussian Mixture Models rank among the best performing frame-wise classification techniques for nonverbal vocalizations which is a reason for us to adopt this technique.

Section 2 presents related work on filler and laughter detection. The data is described in Section 3. We describe our features and method in Section 4 and present our results in Section 5.

2. Related work

We continue to focus on the classes of filled pauses (rather than the broader class of fillers) as the database under study contains filled pauses.

2.1. Filled pause detection

The main characteristic of filled pauses (FPs) that has often been modelled in FP detection is the stability of the oral tract's articulatory configuration during the lengthening of the vowel. Often, researchers use MFCCs and the first two formants [4, 5] as a representation of the articulatory configuration. It is shown in various studies (e.g., [4]) that indeed the standard deviations of F1 and F2 are lower for FPs than for 'normal' speech. Others have used features that aim at modelling the assumed small F0 transition and small spectral envelope deformation [6]. Furthermore, nasality has also been used as a feature as most FPs are

nasalized to a certain extent. Wu and Yan [5] propose to include nasality features based on the first three formants.

In our study, we use MFCCs and the first 2 formants to model spectral properties of FPs. From these features, we derive a spectral similarity measure to capture the spectral stability as a result of the lengthening property of FPs. A nasality feature is added as well.

2.2. Laughter detection

Previous studies on laughter detection have had success by using a set of spectral features such as Perceptual Linear Coding (PLP) or MFCCs [7, 8, 9, 10]. One of the characteristics of laughter that researchers have aimed to capture in features is the occurrence of rhythmic and repetitive laughter calls (i.e. 'laughter syllables') that is less prevalent in speech. This property has been captured by modulation spectrum features [7, 8, 11] that reflect information about syllable rates in speech. GMMs [8], Neural Networks [11], Support Vector Machines [8, 10], Hidden Markov Models [10], and Hidden Conditional Random Fields [10] are among the most popular classification techniques used in laughter detection studies, although not each technique is particularly suitable for frame-wise detection. Knox and colleagues' works [11, 9] on laughter detection specifically focused on frame-wise detection of laughter. They achieved an EER of around 5% on meeting data using MFCC, prosodic features and modulation spectrum features trained in Neural Networks.

In our study, we use MFCCs, pitch and intensity features, formants, and voice quality features to discriminate laughter from other sounds. The first two formants are used since there are indications that F1 and F2 reflect the centralized vowel sounds often encountered in laughter production [12], and that F1 is highly affected by laughter production [13]. We introduce the use of voice quality features for laughter detection to capture the different states of the larynx that are possibly different between laughter and speech [14]. Finally, we investigate whether the relatively simple measure of standard deviation of intensity is able to capture information about the repetitiveness of laughter calls.

3. Data

The data was provided by the organizers of the Computational Paralinguistics Challenge and originates from the SSPNet Vocalisation Corpus (SVC) [15]. Originally, the data is divided into a training, dev, and test set (for which no labels are provided). Because we anticipated the need for an additional separate sub-training set (for example, for training a classifier for fusion), we divided the original training set into two subsets, see Table 1. The training wav files were ordered by name in a list and we attributed the files ordered by uneven numbers of that list to one sub-set and the even numbers to another sub-set. For training our main classifiers, we used the 'uneven' training set.

	Training				Dev	
	'uneven'		'even'			
class	N_{utt}	N_{frames}	N_{utt}	N_{frames}	N_{utt}	N_{frames}
filler	842	41490	865	43544	556	29432
laughter	333	30451	316	28843	225	25750
garbage	1898	796661	1898	794781	1217	492607

Table 1: *Number of frames used in training and testing (N_{utt} is the number of laughter, filler, or garbage utterances).*

We first carried out a short exploration of the database and listened to the data in order to obtain 'a better feeling' for the data. We find that FPs have a mean and median duration of 0.49s (standard deviation of 0.24s) and 0.47s respectively. The shortest duration for an FP is 0.02s and the longest duration is 2.48s. While listening to these extremely short FPs, we observe that some of the shorter FP sounds are in fact lip smack sounds (which were arguably labelled as FPs). For laughter, we find a mean and median duration of 0.91s (standard deviation of 0.68s) and 0.69s respectively. The shortest duration for laughter is 0.15s and the longest 5.1s. We find that many of the shorter laughter sounds are in fact not laughter sounds. In addition, we observe that some laughter calls, that in fact belong to a longer laughter bout, are annotated as separate laughter bouts. This is an observation that we also made in [16] and that has to do with difficulties in defining an appropriate annotation standard for laughter. The longer laughter events sometimes have some speech interspersed with laughter. In sum, one should be aware of these caveats when using the data provided.

4. Features and method

For the extraction of MFCCs `feacalc` [17] were used. For the other features, Praat [18] was used. Each feature is extracted with a steptime of 0.01s. For each frame-wise feature i, we optionally apply functionals, i.e., delta, mean, and standard deviation, that are calculated over a 9-point window (0.09s long) where the ith frame is centred at midpoint.

4.1. Filled pauses characteristics

For FPs, we mainly aim to model their spectral stability and their nasal property through the following features (the number of features is given in brackets, including their label that we use to refer to this feature sets):

Mel-Frequency Cepstrum Coefficients (39, MFCC) MFCCs were extracted with `feacalc` [17]. We extracted 13 MFCCs and their delta and deltadeltas (steptime of 0.01s and analysis window of 0.025s long).

Pitch and intensity (4, PI-FP): Pitch (logarithm of Hz) and intensity features were extracted using a steptime of 0.01s and analysis windows of 0.04s and 0.032s respectively. We used the delta features calculated in a similar way as is done in `feacalc`. For the ith frame, a linear least squares fit was applied to a 9-points analysis window with the ith frame at midpoint. The slopes obtained were used as delta features. Standard deviation was also applied and added as features.

Formants and nasality (14, FORM&NAS): F1 and F2 were extracted (analysis window of 0.025) and their current values, deltas, mean and standard deviations (calculated as described above) were used as features. For nasality, we used a similar energy ratio measure described in [19] where the max energy in the lower range of 0–300Hz is divided by the max energy in the higher range of 300–5500Hz. Further, we measured the peak frequency in the region between 0–800Hz, also suggested by [19]. In addition to their current value, their mean and standard deviations were used.

Spectrum and formant similarity (8, SPECDIST): The Euclidean distances between the current and previous frame of the 39-dimensional MFCC-vector and the 2-dimensional F1F2-vector were calculated and used as features, as well as their delta, mean and standard deviation.

4.2. Laughter characteristics

The following features were used for laughter detection:

Mel-Frequency Cepstrum Coefficients (39, MFCC): These are the exact same MFCCs used for fillers.

Pitch and intensity (8, PI-LAUGH): These are the same pitch and intensity features used for FPs. In addition to the delta and standard deviation features for FPs, we also used the current value and the mean.

Formants (8, FORMANTS): The exact same formant features as used for FPs.

Voice quality (28, VQ): We used voice quality measures based on the Long-Term Averaged Spectrum (LTAS) as described in [20]. In the LTAS (calculated over an analysis window of 0.025s), we measure the max energy in various frequency bands. We denote this as $\text{LTAS}_{0\text{-}2\text{K}}$ which indicates the max energy measured between 0 and 2000Hz. According to [20], the distribution of max energy in the LTAS correlate with perceptions of breathiness, effort, coarseness, and head-chest register. For breathiness, we measured (LTAS_{0-2k} - LTAS_{2k-5k}) – (LTAS_{2k-5k} - LTAS_{5k-8k}) and LTAS_{2k-5k} – LTAS_{5k-8k}. Effort is measured by LTAS_{2k-5k}. Coarseness by LTAS_{0-2k} – LTAS_{2k-5k}. Headchest is measured by (LTAS_{0-2k} - LTAS_{2k-5k}). Further, we included the slope of the LTAS. The current value, delta, mean, and standard deviations of these measurements are used as features.

4.3. Method

GMMs were trained with various number of Gaussian components ranging from 4–256. We trained target (i.e., FPs or laughter) and non-target (i.e., not-FP or not-laughter) GMMs using five iterations of the Expectation Maximization (EM) algorithm. In testing, we obtain frame-wise scores by determining log-likelihood ratios (llr) given the target and non-target GMMs. These frame-wise llrs were then smoothed by applying a median filter for which we tested several sizes ranging from 11–121.

For the combination of several information sources, we apply feature-level fusion by concatenating different features into a higher dimensional feature vector and decision-level fusion by combining the log-likehoods (lls) or log-likelihood ratios of the GMM output. Subsequently, Linear Discriminant Analysis (LDA) is used to train the combinations of lls or llrs. These 'LDA-fusers' are trained with the GMM output of the 'even' subset training data.

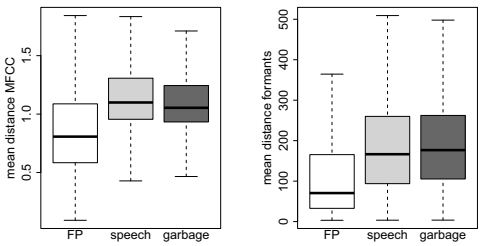
5. Results

5.1. Feature analysis

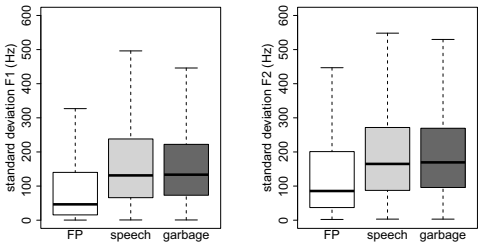
We first inspect whether our intuitions about the features used for FPs and laughter detection are correct and present Box Whisker plots for FPs, laughter, speech, and garbage classes. Since the garbage class also contains silence, we included the speech class which was found by thresholding the sound level and by setting a minimum speech duration of 0.2s.

For FPs, we are mostly interested in the spectral similarity and formant behavior. In Fig. 1 we can observe that indeed the distances between sequencing MFCC and formant vectors are smaller for fillers than for speech or garbage. Similarly, the standard deviation of F1 and F2 are lower for FPs. The distributions of our nasality measures did not appear to differ much from each other.

For laughter, we are interested in the standard deviation of intensity and VQ measures. We can observe in Fig. 2 that the median of standard deviation of intensity for laughter is a bit higher than for speech and garbage but there is still large overlap. Interestingly, for one of the VQ measures, effort, laughter



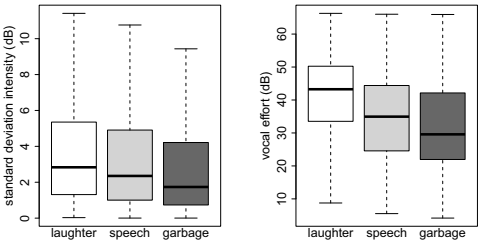
(a) mean distance MFCC (b) mean distance formants



(c) standard deviation F1 (d) standard deviation F2

Figure 1: *Box Whisker plots of various features for FP detection.*

shows higher values, indicating that there is more energy in the higher frequency bands (2k–5k) involved in laughter production.



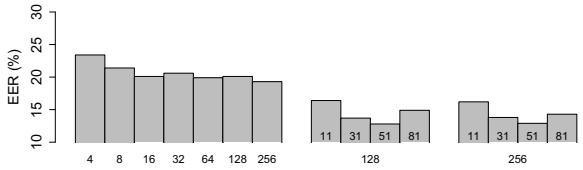
(a) standard deviation intensity (b) vocal effort

Figure 2: *Box Whisker plots of various features for laughter detection.*

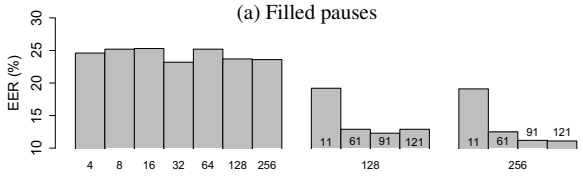
5.2. Classification experiments

We report results of the GMMs in terms of Equal Error Rates (EERs). We first experimented with various number of Gaussians and sizes of median filters trained on all features combined on feature-level (ALL). The results in Fig. 3 show that for both FPs and laughter, the number of Gaussians used matter to a certain extend. Moreover, the use of a median filter improves performance substantially.

Table 2 and 3 report the EERs of the best performing classifiers and the given SVM baseline by feature set. We also trained GMMs for the features provided with the challenge, referred to as COMPARE. For the COMPARE feature set we report the results of the best-performing classifier. To avoid over-specification for our own feature sets, we selected the classifier with the number of Gaussians and median filter size that on overall performed best. For FPs, this yielded a numer of 256 Gaussians and a 51-point median filter. For laughter, a number of 128 Gaussians and a 91-point median filter appeared to work best. We observe that for both FP and laughter detection, the MFCC feature set outperforms all other features, including the ALL set and when combined with the second best performing featureset. The second best performing feature sets are SPECDIST and VQ for FP



(a) Filled pauses



(b) Laughter

Figure 3: *Results of GMMs trained with all features combined on feature-level. For $N_{gauss} = 128$ and 256 we show the EERs when median filters are applied.*

and laughter respectively. As we can observe, median filtering improves the performance substantially with around 7% on average. We further note that our ALL featureset outperforms the challenge's COMPARE and SVM baseline.

Baseline SVM	16.9	
	256 Gaussians	
	No medfilt	51-point medfilt
COMPARE	20.1	13.6
	256 Gaussians	
	No medfilt	51-point medfilt
ALL	19.3	12.9
MFCC	18.4	**10.3**
FORMANTS&NAS	26.9	19.2
SPECDIST	26.0	17.0
PI-FP	31.2	19.0
MFCC + SPECDIST	18.7	10.4

Table 2: *EERs of best perfoming FP detectors (by featureset).*

Baseline SVM	21.2	
	128 Gaussians	
	No medfilt	111-point medfilt
COMPARE	27.7	13.8
	128 Gaussians	
	No medfilt	91-point medfilt
ALL	23.7	12.3
MFCC	21.0	**9.3**
FORMANTS	35.0	23.9
VQ	27.1	17.0
PI-LAUGH	33.8	21.1
MFCC + VQ	23.2	11.7

Table 3: *EERs of best performing laughter detectors (by featureset).*

We attempted to improve the MFCC performance by applying decision-level fusion techniques. An LDA was trained on the log likelihood scores of each target (i.e. FP or laughter) and non-target (i.e. not-FP or not-laughter) GMM of each feature set which yields an 8-dimensional feature vector as input for LDA. Similarly, the log likelihood ratios of each GMM-pair of each feature set were also used as input (4-dimensional feature vector) for LDA. The results are shown in Table 5 and 4. The performances of the LDA-trained classifiers do not outperform the MFCC-trained GMMs. However, the decision-level fusion does give slightly better results compared to a feature-level combination.

FP detection (256 Gaussians)	medfilt	
	-	51p
lls (MFCC, FORM&NAS, SPECDIST, PI-FP)	18.4	12.1
lls (MFCC, SPECDIST)	19.2	11.9
llr (MFCC, FORM&NAS, SPECDIST, PI-FP)	17.9	12.4
llr ((MFCC, SPECDIST)	18.4	11.4

Table 4: *EERs of FP detectors fused with LDA.*

Laughter detection (128 Gaussians)	medfilt	
	-	91p
lls (MFCC, FORMANTS, VQ, PI-LAUGH)	20.0	9.5
lls (MFCC, VQ)	19.3	9.4
llr (MFCC, FORMANTS, VQ, PI-LAUGH)	19.5	9.2
llr (MFCC, VQ)	19.2	9.2

Table 5: *EERs of laughter detectors fused with LDA.*

Finally, we report that we also tried an approach in which we first perform voice activity detection, followed by a normalization of the features over the speech segments obtained. The detection tasks would then become FP vs. speech and laughter vs. speech. This approach however did not yield desirable results and was hence abandoned for the current study.

6. Discussion and conclusion

We developed frame-wise detectors for filled pauses (FPs) and laughter in conversational speech and obtained EERs of 10.3% and 9.3% respectively. The best performance for both FP and laughter detection was obtained with 39 MFCCs and a median filter of 51 and 91 points long. Fusion with other features did not outperform the MFCC performance. Upon inspection of the *unfiltered* llr output, we observed that the variation of sequencing llrs was high and therefore applied median filtering which improved performance substantially. This also suggest that it might be sufficient to produce high scores for those parts in the FP or laughter event that are salient and reliably to detect, and that the filter will smooth out these high scores to neighboring frames.

For future improvements, we inspected the final decisions of the best performing GMMs by thresholding the llrs and compared the segments obtained to the truth labelling. In general we found some very short false positives that could be resolved by setting a minimum duration for FPs or laughter. For fillers, false positives are triggered by clear and long-sounding vowel sounds in for example words such as 'no'. False positives for laughter were usually triggered by breathing sounds. Paradoxically, these errors 'make sense' because the classifiers are trained on detecting exactly these characteristics. One way to tackle these errors is to for example use Hidden Conditional Random Field techniques that can take into account both local and non-local characteristics such that feature sequences and transitions can be modelled more effectively. Finally, we suggest to use these frame-wise based FP and laughter detectors as a basis to move towards real-time detection.

7. Acknowledgements

This work was funded by the EU's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet) and the Dutch national program COMMIT.

8. References

[1] H. H. Clark, *Using Language*. Cambridge University Press, 2005.

[2] J. E. Fox Tree, "Listeners' uses ofum anduh in speech comprehension," *Memory & Cognition*, vol. 29, no. 2, pp. 320–326, 2001.

[3] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting recorder project: Dialog act labeling guide," *ICSI, Berkeley, CA, USA, Tech. Rep. TR-04-002*, 2004.

[4] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken english," *ICASSP*, pp. 4857 – 4860, 2009.

[5] C. H. Wu and G. L. Yan, "Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition," *VLSI Signal Processing*, pp. 91–104, 2004.

[6] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Proceedings of Eurospeech*, 1999, pp. 227–230.

[7] L. S. Kennedy and D. P. Ellis, "Laughter detection in meetings," *ICASSP*, pp. 118 – 121, 2004.

[8] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, pp. 144 – 158, 2007.

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

[10] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," *Perception in multimodal dialogue systems*, pp. 99–110, 2008.

[11] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," *Interspeech*, pp. 2973 – 2976, 2007.

[12] D. P. Szameitat, C. J. Darwin, A. J. Szameitat, D. Wildgruber, A. Sterr, S. Dietrich, and K. Alter, "Formant characteristics of human laughter," in *Proceedings of the Interdisciplinary Workshop The Phonetics of Laughter*, 2007.

[13] D. P. Szameitat, C. J. Darwin, A. J. Szameitat, D. Wildgruber, and K. Alter, "Formant characteristics of human laughter," *Journal of Voice*, vol. 25, pp. 32–37, 2011.

[14] J. H. Esling, "States of the larynx in laughter," in *Proceedings of the Interdisciplinary Workshop The Phonetics of Laughter*, 2007.

[15] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech*, 2013.

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

[17] "SPRACHcore," 2013, http://www1.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html, accessed 18 March 2013.

[18] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[19] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, pp. 225–239, 2004.

[20] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Oto-laryngologica*, vol. 90, pp. 441–451, 1980.