



Laughter and Filler Detection in Naturalistic Audio

Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
The University of Texas at Dallas (UTD), Richardson, Texas, U.S.A.

{lakshmish.kaushik, abhijeet.sangwan, john.hansen}@utdallas.edu

Abstract

Laughter and fillers are common phenomenon in speech, and play an important role in communication. In this study, we present Deep Neural Network (DNN) and Convolutional Neural Network (CNN) based systems to classify non-verbal cues (laughter and fillers) from verbal speech in naturalistic audio. We propose improvements over a deep learning system proposed in [1]. Particularly, we propose a simple method to combine spectral features with pitch information to capture prosodic and spectral cues for filler/laughter. Additionally, we propose using a wider time context for feature extraction so that the time evolution of the spectral and prosodic structure can also be exploited for classification. Furthermore, we propose to use CNN for classification. The new method is evaluated on conversational telephony speech (CTS, drawn from Switchboard and Fisher) data and UT-Opinion corpus. Our results shows that the new system improves the AUC (area under the curve) metric by 8.15% and 11.9% absolute for laughters, and 4.85% and 6.01% absolute for fillers, over the baseline system, for CTS and UT-Opinion data, respectively. Finally, we analyze the results to explain the difference in performance between traditional CTS data and naturalistic audio (UT-Opinion), and identify challenges that need to be addressed to make systems perform better for practical data.

Index Terms: laughter, fillers, deep neural network, convolutional neural network, UT-opinion database, non-verbal cues.

1. Introduction

Non-speech cues such as fillers and laughter play a very important role in human-to-human communication. Fillers are used to hold the floor while planning to allow the necessary time to recollect thoughts, or to prevent the listener from breaking the speaking turn, or to construct thoughts syntactically [2, 3, 4, 6]. Similarly, laughter regulates the flow of interaction, or mitigates the meaning of the preceding utterance. Use of laughter may also signify amusement, happiness, discomfort, scorn or embarrassment [5, 7]. Fillers and laughter also help speakers express emotions and are part of their personality [8, 9]. Hence, automatic detection of laughters and fillers can help in detecting speaker's intentions and emotional state, analyze social behavior of speaker, *etc.*

The problem of detecting laughter and fillers in speech has been investigated previously, and many approaches have been developed to accomplish this task [3]. Research studies such as [8] show that both acoustic and prosodic features help in filler/laughter detection. Other studies have also investigated the use of phonetic patterns and formants, which were shown to perform better than simple cepstrum and pitch based features [9, 10]. A number of machine learning techniques such as Gaussian Mixture Models (GMMs) [12], Maximum Entropy

models, Conditional Random Fields (CRFs), Support Vector Machines (SVMs), Hidden Markov Models (HMMs) [4, 11], Statistical Language Models (SLMs) [4], and Deep Neural Networks (DNNs) [1] have also been investigated for laughter/filler detection task. Among the mentioned machine learning classifiers, DNN based systems seem to provide best results [1].

The mentioned laughter/filler detection systems rely on frame-based processing and decisions are made at frame level. Typical frame length in these classifier systems is between 10ms to 30ms (which is used for feature extraction). However, laughter and fillers are variable length events, and can range from very short (10-20ms) to very long (2-3s). Moreover, in the 10-to-30ms range, laughter/filler can appear very similar to phonemes and a wider time-window may be necessary for better classification. Therefore, in this study, we propose the use of longer time windows for feature extraction.

Laughter and fillers are signals that evolve in time and frequency, and classifiers that can detect and exploit joint time-frequency patterns are likely to model the signal better and deliver superior performance. For example, convolutional neural networks (CNNs) operate simultaneously on the time and frequency dimensions, and have been shown to outperform DNNs in speech recognition tasks [25, 26]. It is reasonable to assume that CNNs should perform better than DNN for laughter/filler detection as well. In this study, we propose a CNN based system for laughter/filler detection.

Naturalistic data collections such as Prof-Life-Log [16, 17] and Apollo Space Missions [30, 31, 33, 32] data are rich in social signals such as filler and laughter. Unlike conversational telephony speech (CTS), these corpora mainly include face-to-face interaction in variety of acoustic environments and stress conditions. It is likely that the challenges posed by naturalistic corpora for laughter/filler detection are different. Additionally, it would be interesting to compare and contrast the performance of laughter/filler detection systems on CTS and naturalistic corpora. In this study, we investigate filler/laughter detection systems on data drawn from Switchboard and Fisher corpora (CTS) and UT-Opinion (a new naturalistic collection described in Sec. 3).

2. Proposed System

The baseline system for this study was developed using [1] (called System (A) in this study). For completeness, we briefly describe the system here. The baseline system uses 141-dimensional (141-d) openSMILE [29] feature set, which includes MFCCs (Mel-filter Cepstral Coefficients), F0, and voicing probabilities. The features are used to train a 3-way DNN (Deep Neural Network) classifier. Particularly, the DNN classifier is trained to distinguish between garbage, filler and laughter. Garbage refers to all non-filler and non-laughter frames (and it

includes speech). The DNN classifier contains two hidden layers (containing 400 neurons each). The DNN output tends to contain fluctuations and post-processing the signal can help improve the overall classification performance. In order to reduce the localized mis-classification, a simple lowpass filter (LPF) is applied on the output state probabilities, namely,

$$H(Z) = \frac{1}{1 - 0.9z^{-1}} \quad (1)$$

The final 3-way decision between laughter, filler and garbage is made by choosing the label corresponding to the highest probability. This constitutes the final output of the system.

In this study, we investigate two variations of the baseline system. In the first variant (called System (B) in this study), we propose a simple feature set for classification. Specifically, we extract 23-dimensional Mel-filter bank energies and 3-d pitch features [34]. Next, we stack 7 frames by splicing 3 frames to left and right of the current frame to form a 182-dimensional feature (7 frames of 26 dimension each). In this manner, the proposed feature set captures both the evolution of spectral and prosodic information in localized time-frequency, which can be exploited for classification. The DNN classifier and post-processing LPF used in System (B) are the same as the one used in System (A).

Research in automatic speech recognition (ASR) systems has shown that deeper networks (*i.e.*, DNNs with 4 or more hidden layers) tend to provide better performance [19, 20, 21, 22, 23]. Therefore, we build another variant (called System (C) in this study), where we use the same feature set and post-processing LPF as in System (B), but train a deeper neural network for classification. Particularly, the DNN in System (C) contains 4 hidden layers with 400 neurons each.

ASR Research has also shown that CNNs typically perform better than DNNs in speech recognition tasks. It is likely that CNNs may outperform DNNs in laughter/filler detection task as well. Therefore, we propose a CNN based classifier. The proposed CNN model (called System (D)) in this study uses the same feature set as Systems (B) and (C). The CNN contains two convolutional and pooling layers, each. The first convolutional layer contains 64 feature-maps of 5x5 dimensions, followed by a pooling layer of 1x3 dimensions. The second layer contains 64 feature-maps of 2x2 dimensions, followed by a pooling layer of 1x1 dimension. The output of the second pooling layer is fed into a neural network with two hidden layer of 400 neurons each. The output label probabilities from the CNN classifier are post-processed using the same LPF as in Systems (A), (B) and (C).

In this study, we used the Kaldi toolkit for extracting mel-filterbank and pitch features [18]. Additionally, the PDNN toolkit was used for training and evaluating the DNN and CNN classifiers [27].

3. Databases

3.1. Conversational Telephony Speech (CTS)

In this study, we use audio files from switchboard and fisher corpora to form the CTS evaluation set. We created two CTS datasets: (i) the 5K dataset which contained 4000 and 1000 training and evaluation files, respectively, and (ii) the 30K dataset which contained 25000 and 5000 training and evaluation files, respectively. The 5K dataset facilitated a quicker experiment cycle which allowed us to try various configurations and fine tune the system.

Table 1: Training and evaluation dataset used in this study. (G: Garbage, L: Laughter, F: Filler)

Database Variation	# Utterances	Duration (Hrs)	Duration Split (%)			No. of Events	
			G	F	L	F	L
SWB+Fisher Small (5K)	Train: 5000	2.0	60	20	20	3012	1597
	Eval: 1000	0.5	58	17	25	630	493
SWB+Fisher Big (30K)	Train: 30000	13.0	58	17	25	15259	12731
	Eval: 5000	2.5	60	19	21	3391	2232
UT-Opinion	Eval: 152	1.5	93	5.7	1.3	981	164

In both the 5K and 30K datasets, speech recognition models were used to perform phone-level forced alignment (which included all phonemes, silence, laughter and fillers). Phoneme and silence frames were assigned to garbage label. Laughter and filler frames were assigned to laughter and filler labels, respectively. Table 1 shows the dataset details. Particularly, the duration of 5K and 30K train and evaluation sets are shown. Additionally, the duration percentage split between laughter, filler and garbage labels for the 5K and 30K datasets are also shown. Finally, the total number of unique laughter and filler events in the datasets can also be seen. In general, the dataset used in this study is larger than the one used in [1].

3.2. UT-Opinion

UT-Opinion is a new corpus developed by us to study sentiment and opinion in naturalistic conversations [14, 15]. In UT-Opinion, each subject is interviewed (face-to-face conversation style) and is asked to share his/her opinion on 10 different topics. Subjects are interviewed at various locations on the University of Texas at Dallas (UTD) campus including classrooms, hallways, office rooms, library, gymnasium and street. The subjects include students, staff and faculty members of both genders. Additionally, both native and non-native speakers are part of the collection. Altogether, data from 120 subjects has been collected so far resulting in 1200 evaluation audio files. From this data, we have selected a subset of 1.45 hours of data (152 audio files). Files that contained at-least 3 fillers and/or 1 laughter were selected for evaluation in this study. Altogether, the 152 files contained 45 unique speakers, 164 laughter events and 981 filler instances. The details of the dataset are also shown in Table 1.

UT-Opinion is an interesting corpus of laughter/filler evaluation because the participants are engaged in face-to-face interaction (which is different from traditional CTS corpora). The availability of a visual channel between speakers may have an impact on non-verbal signals such as filler and laughter. Additionally, the interaction occurs in a variety of natural settings where the acoustic environment can range from clean to noisy (which makes the task challenging from a robustness perspective). The speakers in UT-Opinion belong to different nationalities/cultures which is potentially a factor contributing to variability in filler and laughter signal.

4. Results and Discussions

4.1. System Evaluation: CTS datasets

In this first experiment, we compare the performance of baseline System (A) against the proposed Systems (B), (C) and (D). All 4 systems are trained and evaluated using the 5K train and evaluation datasets, respectively. In Table 2, the performance accuracy of each system is shown in terms of AUC (area under the curve) metric [1]. The performance of each system with and without the post-processing LPF (low pass filter) is also shown.

Table 2: AUC scores laughter, filler and garbage detection in CTS 5k, 30K and UT-Opinion for Systems (A), (B), (C), (D) and (E).

System	Event	Baseline (A)	(A) + LPF	DNN (B)	(B) + LPF	Deeper DNN (C)	(C) + LPF	CNN (D)	(D) + LPF	CNN (E)	(E) + LPF
Switchboard + Fisher (AUC)	Garbage	86.7	88.75	90	91.92	90	91.16	91.8	93.2	94.2	95.07
	Filler	92.9	94.36	95.6	96.21	95.4	95.97	96.2	96.93	97.4	97.75
	Laughter	88.3	91.96	91.7	94.41	90.9	93.6	92.4	94.85	94.5	96.45
UT-Opinion (AUC)	Garbage	74.02	76.64	79.25	81.59	78.97	80.77	79.68	82.42	78.81	81.72
	Filler	79.42	81.808	82.51	83.93	82.24	83.32	83.63	85.49	83.4	85.43
	Laughter	72.03	76.16	76.67	80.49	76.26	79.4	78.09	81.79	80.34	83.9

The AUC for laughter and filler detection due to the baseline system (A) is 88.3 and 92.9, respectively. The AUC performance numbers for the baseline system are comparable to those reported in [1] (although it should be noted that the two studies use different evaluation datasets). Furthermore, using LPF for post-processing improves the AUC numbers for both filler and laughter. A similar observation was made in [1].

The AUC for laughter and filler detection due to System (B) is 91.7 and 95.6, respectively. Since the only difference between Systems (A) and (B) are the input features, the proposed feature set outperforms the 141-dimensional openSMILE feature set in this study. One reason for this could be that the proposed feature set (containing multiple concatenated frames of Mel-filter bank energies and pitch features) captures a wider time-context which facilitates better classification. Additionally, it is also possible that the spectral and prosodical information provided by filter bank energies and pitch are sufficient for laughter/filler detection.

For System (C), the laughter and filler AUC numbers are 93.6 and 95.97, respectively. Hence, the performance for (C) is inferior to (B). Since the only difference between Systems (B) and (C) is that (B) uses a deeper neural network, it seems like the use of more hidden layers did not help in improving performance accuracy.

Furthermore, for System (D), the laughter and filler AUC numbers are 92.4 and 96.2, respectively. Hence, the performance for (D) is better than (B). This result suggests that CNN may indeed be a better classifier for laughter/filler detection as opposed to DNN. This result is consistent with the experience of DNN and CNN classifiers in the speech recognition community.

4.2. System Evaluation: UT-Opinion

In the next experiment, we trained System (D) with the larger 30K dataset. The purpose of the experiment was to build a larger system for UT-Opinion evaluation. In this study, we call this System (E). The laughter and filler AUC performance for System (E) on 30K dataset are also shown in Table 2.

The laughter and filler detection performance in terms of AUC for UT-Opinion evaluation dataset for Systems (A), (B), (C), (D), and (E) are shown in Table 2. As seen for CTS evaluation, the performance of System (B) is better than System (A), System (C) does not improve performance over System (B), and System (D) outperforms System (B). The use of LPF post-processing step always helps in improving system performance.

Interestingly, the performance of System (E) is not consistently better than (D). While laughter AUC number for (E) is better than (D), the filler AUC number is marginally inferior. Since the only difference between (D) and (E) is the size of the training set, it is possible that more training data for fillers is not

useful for learning. However, more training data for laughter seems to help. One reason for this result could be that the fillers in UT-Opinion (evaluation) are somewhat different from those in the CTS (training) datasets, resulting in mismatch (and increasing the training data does not reduce the mismatch). More investigation is required to better understand this phenomenon.

Overall, the system performance suggests that CNNs outperform DNNs, LPF based post-processing is useful to obtain improved results, and simple spectral and pitch based features may be sufficient for laughter/filler detection.

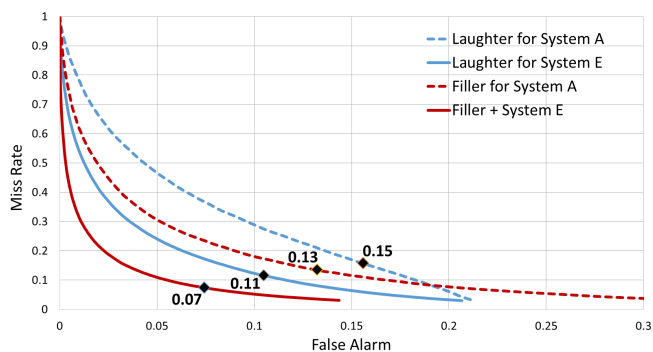


Figure 1: DET Curves of Laughter and Filler for Systems A+LPF and E+LPF for CTS 30K data.

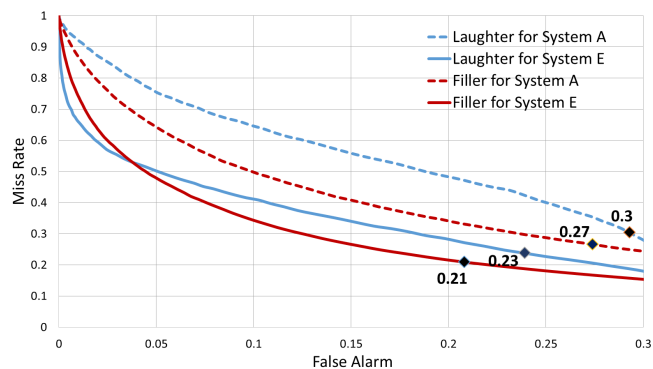


Figure 2: DET Curves of Laughter and Filler for Systems A+LPF and E+LPF for UT-Opinion data.

4.3. Detection Error Tradeoff (DET) Curves

Figures 1 and 2 show the DET curves for CTS 30K and UT-Opinion evaluation datasets, respectively. In order to plot the DET curves, we varied the threshold for classification from 0 to 1, and the frame was classified as laughter or filler if the probability of laughter or filler exceeded the threshold, respectively.

From Figures 1 and 2, it is seen that System (E) outperforms (A) across all operating points for both laughter and filler detection. Additionally, filler detection is more accurate than laughter detection across all operating points for the CTS 30K evaluation dataset. However, in the case of UT-Opinion evaluation dataset, the performance of laughter and filler detection is somewhat comparable. While the performance of both laughter and filler detection decreases when moving from CTS to UT-Opinion, it seems like filler detection performance is more negatively impacted than laughter detection. This observation is consistent with what was seen in the AUC numbers in table 2.

Table 3: Frame level confusion matrix for the Systems (A)+LPF and (E)+LPF. (G: Garbage, L: Laughter, F: Filler)

SWB + Fisher Database						UT-Opinion Database									
System A (%)			System E (%)			System A (%)			System E (%)						
G	F	L	G	F	L	G	F	L	G	F	L				
G	79.84	10.27	9.88	G	90.54	3.56	5.90	G	65.83	18.90	15.26	G	86.77	4.18	9.04
F	14.17	78.57	7.25	F	12.78	82.62	4.59	F	21.13	71.42	7.44	F	22.69	70.76	6.55
L	30.10	8.95	60.95	L	23.65	3.05	73.30	L	44.86	7.4	47.73	L	33.80	1.08	65.11

4.4. Analysis

The classification confusion matrix for laughter and filler detection is shown for UT-Opinion and CTS 30K datasets in Table 3. The tables suggest that the biggest source of errors is laughter getting mis-classified as garbage followed by filler getting mis-classified as garbage. On the other hand, the smallest source of errors is laughter getting mis-classified as filler and vice-versa. Comparing System (A) and (E) performance suggests that (E) reduces both filler vs. garbage and laughter vs. garbage confusions. However, laughter and filler getting mis-classified as garbage remain the major source of errors and could be the subject of future studies.

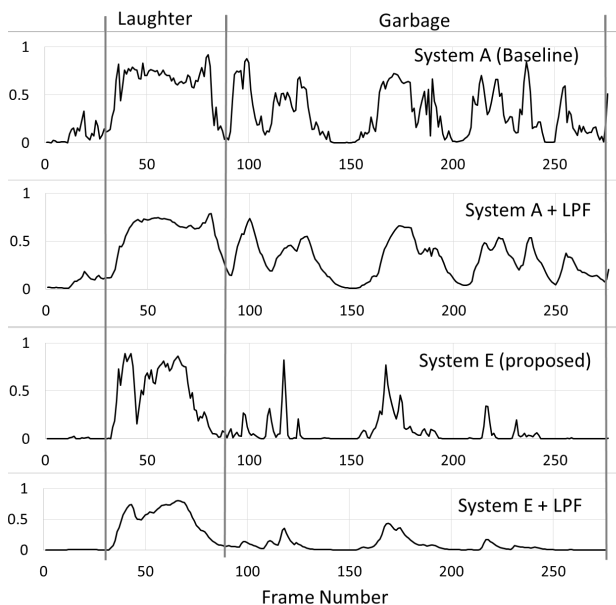


Figure 3: Improved laughter detection due to CNN and LPF over baseline DNN system.

Figure 3 shows the sample output for Systems (A), (A)+LPF, (E), and (E)+LPF for a 3 second audio clip containing laughter. System (A) output contains a lot of fluctuations which causes a number of false-positives (garbage detected as laugh-

ter). Using LPF has beneficial effect as it reduces the fluctuations (see output of (A)+LPF). A similar analysis highlighting the benefits of LPF was shown in [1]. Additionally, the output of System (E) can also be seen in the figure. It can be observed that the CNN output is better able to separate laughter vs. garbage, and causes few false-positives. System (E) also captures a wider time-context in its feature set which may also help in providing a smoother output. Additionally, the CNN system also benefits from LPF as the number of potential false-positives are further reduced.

In an effort to explain the difference in performance on the UT-Opinion and CTS datasets, we compute and compare the laughter and filler duration distributions for the two datasets. The distributions are shown in Fig. 4. From the figure, it can be seen that the average laughter and filler duration for UT-Opinion is greater than CTS corpora. It is likely that the difference in filler and laughter duration is not responsible for the poorer performance of laughter and filler detection on UT-Opinion (intuitively we would expect same or better performance for longer duration laughter and filler events). This suggests that the source of mismatch is perhaps spectral or prosodic. This is possible as UT-Opinion has greater diversity of acoustic backgrounds, accents/dialects, etc. More investigation is required to understand the reasons for the performance gap better.

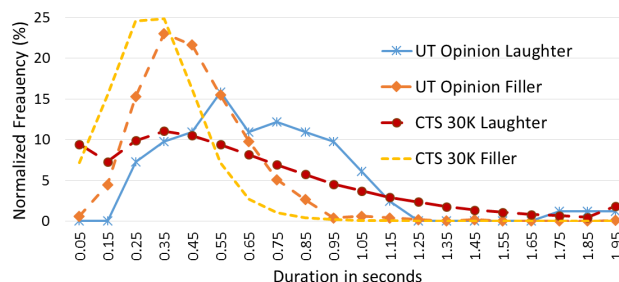


Figure 4: Duration histograms for laughter and filler for both UT-Opinion and CTS 30K datasets.

5. Conclusions

In this study, a new system for filler and laughter detection has been proposed. The new system utilizes a simple feature set which combines spectral (Mel-filter bank energies) and prosodic (pitch) information. Additionally, the feature set also captures a wider time-context which was shown to be useful for classification. The new system uses Convolutional Neural Networks (CNNs) followed by simple low-pass filtering (LPF) for classification. The proposed system was evaluated on Conversational Telephony Speech (CTS) and UT-Opinion datasets, and was shown to outperform a Deep Neural Network (DNN) based system. Furthermore, analysis of the system revealed that the mis-classification of laughter and filler as garbage is a major source of errors and a topic of future research. Finally, more research is required to improve the performance of laughter and filler detection systems on naturalistic audio data (such as UT-Opinion).

6. Acknowledgement

This material is based upon work supported in part by AFRL under contract FA8750-12-1-0188, by NSF under Grant 1218159, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen

7. References

- [1] R. Gupta, K. Audhkhasi, S. Lee and S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking.", *Interspeech'13*, pp. 173-177, 2013.
- [2] M. Argyle, V. Salter, H. Nicholson, M. Williams and P. Burgess, "The communication of inferior and superior attitudes by verbal and nonverbal signals," *British journal of social and clinical psychology*, vol. 9, no. 3, pp. 222-231, 2011.
- [3] J. Morreall, "Taking laughter seriously", State University of New York Press, 1983.
- [4] H. Salamin, A. Polychroniou and A. Vinciarelli, "Automatic Detection of Laughter and Fillers in Spontaneous Mobile Phone Conversations.", *IEEE International Conference on Systems, Man and Cybernetics*, pp. 4282-4287, 2013.
- [5] A. Vinciarelli, M. Pantic and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743-1759, 2009.
- [6] I. Poggi and F. D. Errico, "Social Signals: a framework in terms of goals and beliefs," *Cognitive Processing*, vol. 13, no. 2, pp. 427-445, 2012.
- [7] P. Brunet and R. Cowie, "Towards a conceptual framework of research on social signal processing," *Journal of Multimodal User Interfaces*, vol. 6, no. 3-4, pp. 101-115, 2012.
- [8] 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, Jan. 2004
- [9] J. E. Fox Tree, "Listeners uses of um and uh in speech comprehension," *Memory and Cognition*, vol. 29, no. 2, pp. 320-326, 2001.
- [10] 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.
- [11] 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.
- [12] T. F. Krikke and K. P. Truong, "Detection of nonverbal vocalizations using Gaussian Mixture Models: looking for fillers and laughter in conversational speech," *Interspeech*, pp. 1637-167, 2013.
- [13] B. Schuller, S. Steidl, A. Batliner et. al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *Interspeech*, 2013.
- [14] L. Kaushik, A. Sangwan and J.H.L. Hansen, "Sentiment extraction from natural audio streams," *ICASSP*, May 2013.
- [15] L. Kaushik, A. Sangwan and J.H.L. Hansen, "Automatic sentiment extraction from YouTube videos," *Automatic Speech Recognition and Understanding (ASRU)*, 2013.
- [16] A. Ziaei, A. Sangwan and J. H. L. Hansen, "Prof-Life-Log: Audio Environment Detection for Naturalistic Audio Streams," *Interspeech*, 2012.
- [17] A. Ziaei, A. Sangwan and J. H. L. Hansen, "Prof-Life-Log: Personal interaction analysis for naturalistic audio streams," *ICASSP*, 2013.
- [18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," *ASRU*, 2011.
- [19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527-1554, 2006.
- [20] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, 29(6), 82-97, 2012.
- [21] G. E. Dahl, Y. Dong, L. Deng and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 30-42, 2012.
- [22] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," *Interspeech*, 2011.
- [23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533-536, 1986.
- [24] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in NIPS*, pp. 1096-1104, 2009.
- [25] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition.", *Interspeech*, 2013.
- [26] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying Convolutional Neural Network Concepts to Hybrid NN-HMM Model for Speech Recognition," *ICASSP*, 2012.
- [27] Y. Miao, "Kaldi+PDNN: building DNN-based ASR systems with Kaldi and PDNN.", *arXiv preprint arXiv:1401.6984*, 2014.
- [28] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance.", *NIST*, Gaithersburg Maryland, 1997
- [29] F. Eyben, M. Wollmer and B. Schuller, B. "Opensmile: the Munich versatile and fast open-source audio feature extractor," *The International Conference on Multimedia*, pp. 1459-1462, 2010.
- [30] A. Sangwan, L. Kaushik, C. Yu, J.H.L. Hansen, and D. W. Oard, "Houston, we have a solution: using NASA apollo program to advance speech and language processing technology," *Interspeech*, 2013.
- [31] A. Ziaei, L. Kaushik, A. Sangwan, J.H.L. Hansen, and D. W. Oard, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions.", *Interspeech*, 2014.
- [32] D. Oard, A. Sangwan, and J.H.L. Hansen. "Reconstruction of apollo mission control center activity." *First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH)*, 2013.
- [33] C. Yu, J.H.L. Hansen, and D.W. Oard. "Houston, We have a solution: A Case Study of the Analysis of Astronaut Speech during NASA Apollo 11 for Long-Term Speaker Modeling." *Interspeech*, 2013.
- [34] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," *ICASSP*, 2014.