Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 72–92

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Detecting paralinguistic events in audio stream using context in features and probabilistic decisions☆

Rahul Gupta[a,*], Kartik Audhkhasi[b], Sungbok Lee[a], Shrikanth Narayanan[a]

[a] *Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Avenue, Los Angeles, CA 90089, USA*
[b] *IBM Thomas J Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA*

Received 28 November 2014; received in revised form 24 June 2015; accepted 7 August 2015
Available online 11 September 2015

Abstract

Non-verbal communication involves encoding, transmission and decoding of non-lexical cues and is realized using vocal (e.g. prosody) or visual (e.g. gaze, body language) channels during conversation. These cues perform the function of maintaining conversational flow, expressing emotions, and marking personality and interpersonal attitude. In particular, non-verbal cues in speech such as paralanguage and non-verbal vocal events (e.g. laughters, sighs, cries) are used to nuance meaning and convey emotions, mood and attitude. For instance, laughters are associated with affective expressions while fillers (e.g. um, ah, um) are used to hold floor during a conversation. In this paper we present an automatic non-verbal vocal events detection system focusing on the detect of laughter and fillers. We extend our system presented during Interspeech 2013 Social Signals Sub-challenge (that was the winning entry in the challenge) for frame-wise event detection and test several schemes for incorporating local context during detection. Specifically, we incorporate context at two separate levels in our system: (i) the raw frame-wise features and, (ii) the output decisions. Furthermore, our system processes the output probabilities based on a few heuristic rules in order to reduce erroneous frame-based predictions. Our overall system achieves an Area Under the Receiver Operating Characteristics curve of 95.3% for detecting laughters and 90.4% for fillers on the test set drawn from the data specifications of the Interspeech 2013 Social Signals Sub-challenge. We perform further analysis to understand the interrelation between the features and obtained results. Specifically, we conduct a feature sensitivity analysis and correlate it with each feature's stand alone performance. The observations suggest that the trained system is more sensitive to a feature carrying higher discriminability with implications towards a better system design.

Keywords: Paralinguistic event; Laughter; Filler; Probability smoothing; Probability masking

1. Introduction

Non-verbal communication involves sending and receiving non-lexical cues amongst people. Modalities for transmitting non-verbal cues include body language, eye gaze and non-verbal vocalizations. Non-verbal communication is hypothesized to represent two-thirds of all communication (Hogan and Stubbs, 2003) and its primary functions

☆ This paper has been recommended for acceptance by Björn Wolfgang Schuller.
[*] Corresponding author. Tel.: +1 2135871103.
E-mail address: guptarah@usc.edu (R. Gupta).

http://dx.doi.org/10.1016/j.csl.2015.08.003
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

Table 1
Statistics of laughter and filler annotations in the SVC corpus.

Event	Total number of segments	Statistics over the segment lengths (in milliseconds)		
		Mean	Standard deviation	Range
Laughter	1158	943	703	2–5080
Filler	2988	502	262	1–5570

include reflecting attitude and emotions (Argyle et al., 1970; Mehrabian and Ferris, 1967; Halberstadt, 1986), assisting dialog process (Bavelas and Chovil, in press; Johannesen, 1971) as well as expressing personality (Isbister and Nass, 2000; Cunningham, 1977). Studies suggest that non-verbal communication is a complex encoding-decoding process (Zuckerman et al., 1975; Lanzetta and Kleck, 1970). Encoding relates to the generation of non-verbal cues, usually in parallel with verbal communication and decoding involves interpretation of these cues (Argyle, 1972; O'sullivan et al., 1994). Studies broadly classify non-verbal communication into two categories, visual and vocal (Streeck and Knapp, 1992; Poyatos, 1992). Visual cues include communication through body gestures, touch and body distance (Ruesch and Kees, 1956) and vocal cues comprise paralanguage (e.g., voice quality, loudness) and non-verbal vocalizations (e.g., laughters, sighs, fillers) (Schuller et al., 2008; Bowers et al., 1993). Both these channels of non-verbal communication have been extensively studied and the literature suggests their relationship to varied phenomena and constructs including language development (Harris et al., 1986), child growth (Mundy et al., 1986; Curcio, 1978), relationship satisfaction (Kahn, 1970; Boland and Follingstad, 1987) and psychotherapy process (Gupta et al., 2014). This extension of non-verbal communications research beyond understanding their primary functions reflects their significance in interaction.

Our focus in this work is on non-verbal vocalizations (NVVs) in speech. Previous research links various forms of non-verbal vocalizations such as laughters, sighs and cries to emotion (Goodwin et al., 2009; Gupta et al., 2012), relief (Soltysik and Jelen, 2005; Vlemincx et al., 2010) and evolution (Furlow, 1997). The importance of each of these non-verbal vocalizations is highlighted by the role they play in human expression. Therefore a quantitative understanding of their production and perception can have a significant impact on both behavioral analysis and behavioral technology development. In this paper, we aim to contribute to the analysis of these non-verbal vocalizations by developing a system for detection of non-verbal events in spontaneous speech.

Several previous works have proposed detection methods for NVVs. Kennedy and Ellis (2004) demonstrated the efficacy of using window-wise low level descriptors from speech (Cortes and Vapnik, 1995) in detecting laughters in meetings. Truong and Van Leeuwen (2005) investigated perceptual linear prediction (PLP) and acoustic prosodic features for NVV detection using Gaussian mixture models. Várallyay et al. (2004) performed acoustic analysis of infant cries for early detection of hearing disorders. Schuller et al. (2008) presented static and dynamic modeling approach for recognition of non-verbal events such as breathing and laughter in conversational speech. In particular, the Interspeech 2013 Social Signals Sub-challenge (Schuller et al., 2013) led to several investigations (Kaya et al., 2013; Pammi and Chetouani, 2013; Krikke and Truong, 2013; Brueckner and Schulter, 2014; An et al., 2013) on frame-wise detection of two specific non-verbal events: laughters and fillers. Building upon on our efforts (Gupta et al., 2013) on the same challenge dataset (Salamin et al., 2013) (that was the winning entry in the challenge), in this paper we perform further analysis and experiments. Previous works in this research field have primarily focused on local characteristics and our approach investigates the benefits of considering context during the frame-wise prediction. Our methods are inspired from the fact that the non-verbal events occur over longer segments (and hence analysis frames). The temporal characteristics of these events has been investigated in studies such as (Mowrer et al., 1987; Bachorowski et al., 2001; Candea et al., 2005). These studies reveal interesting patterns such as a positive correlation between duration of laughter (Mowrer et al., 1987) and number of intensity peaks and similarity in duration of fillers across languages (Candea et al., 2005). Bachorowski et al. (2001) went further into the details of laughter types (e.g. voiced vs unvoiced) and their relation to laughter durations. More studies on laughter and filler duration and its relation to their acoustic structures can be found in (Vettin and Todt, 2004; Sundaram and Narayanan, 2007; Vasilescu et al., 2005). As statistics (presented later in Table 1) on our database of interest also show that laughters and fillers exist over multiple analysis frames, we hypothesize that information from neighboring frames can be utilized to reduce the uncertainty associated with the current frame.

Table 2
Statistics of data splits used as training, development and testing set.

Count	Dataset			Total
	Training	Development	Testing	
Clips	1583	500	680	2763
Laughters Segments	649	225	284	1158
Fillers Segments	1710	556	722	2988

Given that the target events are temporally contiguous, one can use many of the available sequential classifiers for the problem of interest. Potential techniques include Markov models (Rabiner and Juang, 1986), recurrent neural networks (Funahashi and Nakamura, 1993) and linear chain conditional random fields (Lafferty et al., 2001). For instance, Cai et al. (2003) used Hidden Markov Models (HMM) for detection of sound effects like laughter and applause from audio signals. This approach is similar to methods used in continuous speech recognition (Rabiner and Juang, 1986) with a generative model (the HMM), which may not be as optimal as other discriminative methods in event detection problems (Tu, 2005). Brueckner and Schuller (2013) used a hierarchical neural network for detecting audio events. This model initially provides a frame-level prediction using low level features and then another neural network is used to combine predictions from multiple frames to provide a final frame-level prediction. Other studies (Kaya et al., 2013; Piccardi, 2004) have also used similar prediction ad-hoc filtering methods (median filtering, Gaussian smoothing) to incorporate context in similar sequence classification problems. Most of these works have focused on performance driven approaches towards design of the detection systems but fail to provide a thorough model and feature analysis. In this work, we explore and analyze new architectures to incorporate context at the two levels of (i) frame-wise acoustic features and, (ii) frame-wise probabilistic decisions obtained from the features as a continuation of previous effort (Gupta et al., 2013). Through our analysis in this paper, we aim to understand the relation of laughters and fillers to the low level features as well as the temporal characteristics of these events. We focus on aspects such as using contextual features during classification and incorporating context in frame-wise outputs (termed as ‘smoothing’ and ‘masking’ operations). Our final system achieves an area under the receiver operating characteristics (ROC) value of 95.3% for laughters and 90.4% for fillers on a held out test set. We also present an analysis on the role of each feature used during detection and its impact on the final outcome.

The paper is organized as follows: Section 2 provides the database description and statistics and Section 3 lists the set of used features. We present our core NVV detection schemes inclusive of smoothing and masking in Section 4. Section 5 provides feature analysis and conclusion is presented in Section 6.

2. Database

We use the SSPNet Vocalization corpus (SVC) (Salamin et al., 2013) for the experiments in this paper. This data was used as the benchmark during the Interspeech challenge and provides a platform for comparison of various algorithmic methods (Kaya et al., 2013; Pammi and Chetouani, 2013; Krikke and Truong, 2013; Brueckner and Schulter, 2014; An et al., 2013) The dataset consists of 2763 audio clips, each 11 seconds long. Each of these clips have at least one filler or laughter event in between 1.5 s and 9.5 s. These clips are extracted from 60 phone calls involving 120 subjects (57 male, 63 female) containing spontaneous conversation. The pair of participants in each call perform a winter survival task which involves identifying an entry from a predefined list consisting of objects useful in a polar environment. The conversation was recorded on cellular phones (model Nokia N900) at the two ends. There was no overlap in speech as the recordings were made separately for the speakers involved. The audio files are manually annotated (single annotator) for laughter and filler. We list the statistics for laughter and filler events over the entire database in Table 1. For more details on the dataset please refer to (Salamin et al., 2013; Schuller et al., 2013).

For modeling and evaluating the frame-wise detection, we use the training, development and testing splits as defined during the Interspeech challenge (Schuller et al., 2013). The speaker information per clip is not available but the training, development and testing sets contain non-overlapping set of speakers (training: speakers 1-70, development: speakers 71-90, testing: speakers 91-120). Non-overlapping set of speakers allows for speaker-independent evaluation. Table 2

Table 3
Set of features extracted per frame.

Prosodic features	Voicing probability
	Harmonic to noise ratio
	Fundamental frequency (F0)
	Zero crossing rate
	Log intensity
MFCC	12 coefficients

shows the counts of clips, laughter segments and filler segments in each data split. In the next sections, we list the set of features extracted per audio clip followed by our NVV detection scheme.

3. Feature extraction

We use an assembly of prosodic features and mel-frequency cepstral coefficients (MFCCs) extracted at a frame-rate of 10 milliseconds using OpenSMILE (Eyben et al., 1462). This set of features is same as the one used in the 2013 Interspeech challenge (Schuller et al., 2013) and has been previously used in several other classification and detection experiments involving emotion, depression and child behavior characterization (Liu et al., 2006; Pierre-Yves, 2003; Bone et al., in press; Bone et al., 2013; Gupta et al., 2012). Studies (Kennedy and Ellis, 2004; Truong and Van Leeuwen, 2005) have shown the relation of speech prosody and spectral characteristics to similar non-verbal events. Note that this list of features, while large, is by no means exhaustive and new features for laughter and filler detection have been proposed in several other works (An et al., 2013; Wagner et al., 2013). As we focus on the system development aspect of event detection in this paper, we work with this smaller representative set of features provided during the Interspeech challenge. For further improvement and specific feature analysis, the proposed system can definitely be augmented with the new sets of features in future. The features used in this paper are listed in Table 3. We z-normalize these features per file before subsequent system training. The feature means and variances for normalization are calculated over the entire duration of the file.

4. Event detection scheme

We use the aforementioned feature set to train our frame-wise detection scheme. We test the effect of incorporating context at various stages in detecting an event $E \in$ {laughter, filler}. As these events usually last over multiple frames, we hypothesize that proximal information may help in detection of these events. Given the rare occurrence of these events, the Interspeech challenge (Schuller et al., 2013) adopted area under the Receiver Operating Characteristics (ROC) curve as the evaluation metric for laughter and filler detection. We use the same metric and explore several prediction architectures to maximize it. We develop a three step sequential algorithm which involves:

(i) Predicting event probabilities based on the speech features. We investigate the effect of incorporating contextual features and compare it with a context independent baseline.
(ii) Incorporating context in probabilistic frame-wise outputs obtained from the previous step.
(iii) ‘Masking’ the smoothed probabilities based on heuristics based rules.

Fig. 1 provides a block diagram representation of our methods and constituent experiments. We describe each of these steps below.

4.1. *Predicting event probabilities based on the speech features*

We initiate our system training procedure with a model that uses the speech features and outputs the frame-wise probabilities of an event E. First, we present a model that takes no context into account and serves as a baseline. Next,

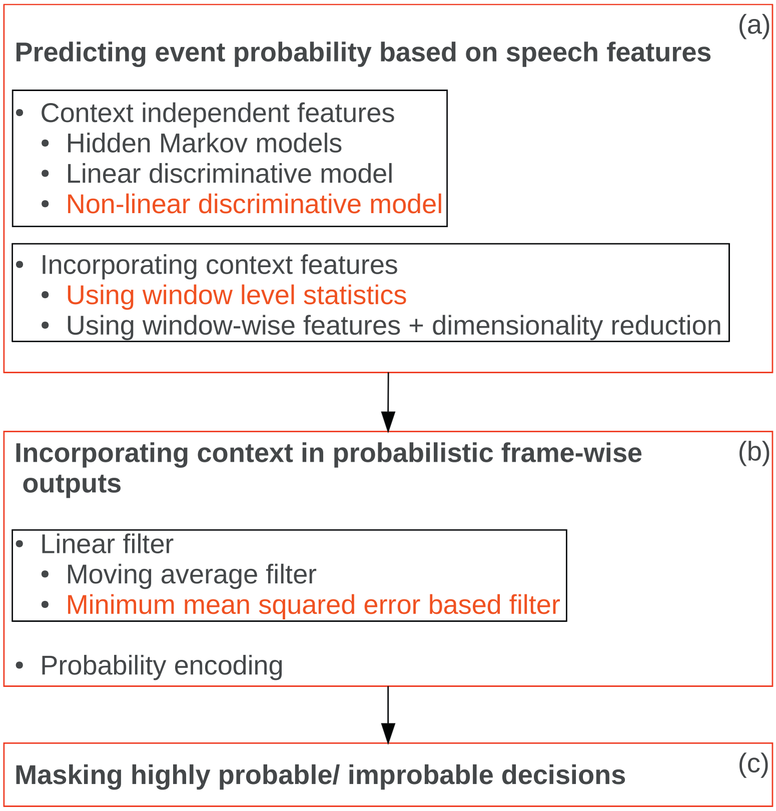
(a)
Predicting event probability based on speech features
Context independent features
Hidden Markov models
Linear discriminative model
Non-linear discriminative model
Incorporating context features
Using window level statistics
Using window-wise features + dimensionality reduction
(b)
Incorporating context in probabilistic frame-wise outputs
Linear filter
Moving average filter
Minimum mean squared error based filter
Probability encoding
(c)
Masking highly probable/ improbable decisions

Fig. 1. Block diagram representing each processing step performed during detection. Contents in gray boxes show experimental methods used with the best method (as determined during the system development) marked in red. We retain the best method at each step for subsequent processing step. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we augment the feature set by introducing certain contextual features. We next describe the baseline scheme followed by the incorporation of contextual features.

4.1.1. Baseline: frame-wise classification with context independent features

In this classification scheme, we obtain the event probability exclusively based on the 17 features extracted per frame, as listed in Table 3. Note that these features represent the acoustic characteristics of only the analysis frame under consideration and contain no information about the feature values from neighboring frames or the feature dynamics. We train models to assign each audio frame as belonging to a target event or as a garbage frame based on the acoustic features. We experiment with several classification architectures and model our class boundaries using (i) a generative Hidden Markov Model (Rabiner and Juang, 1986) (ii) a linear discriminative classifier (used as baseline in (Schuller et al., 2013)) and (iii) a non-linear discriminative classifier to assess the nature of separability of event classes in the feature space. We represent the column vector of features for the nth frame as $\mathbf{x}_n$ and the corresponding probability obtained for an event E as $u_E(n)$. We describe the training methodology for each of the models below.

(i) Classification with Hidden Markov Model (HMM): In this scheme, we train an HMM model using the Kaldi toolkit (Povey et al., 2011). We train monophone models using the Viterbi-EM algorithm (Forney, 1973) for each of the laughter, filler and garbage events from the training set. We then decode the development and the test sets using a unigram language model (LM) assuming equal LM weights for the three events (to account for class imbalance in the training set, as majority of frames belong to the garbage class). During decoding, we obtain the state occupancy probabilities of each frame to be belonging to the laughter, filler or garbage HMM. We use these probabilities from each of these HMMs as our laughter, filler and garbage probabilities.

(ii) Classification with a linear discriminative classifier: We determine linear class boundaries using the Support Vector Machine (SVM) classifier (this model was also used in the Interspeech challenge (Schuller et al., 2013)). We

Table 4
AUC for prediction using context independent features with HMM, linear and non-linear classifiers.

System used	AUC (in %)			
	Development set		Testing set	
	Laughter	Filler	Laughter	Filler
Chance	50.0	50.0	50.0	50.0
Hidden Markov Models	74.2	79.3	71.3	78.0
Linear discriminative classifier	77.0	81.2	73.8	79.1
Non-linear discriminative classifier	83.7	83.8	80.9	82.3

train a multi-class SVM classifier over the three classes with pair-wise boundaries. We obtain the class probabilities for each frame by fitting logistic regression models to data-point distances from the decision boundaries. In order to prevent class bias due to unbalanced representative data from each class, we downsample the 'garbage' frames (frames not belonging to either of the events) by a factor of 20 during training, as suggested in the challenge paper (Schuller et al., 2013). We use a linear kernel and the slack term weight is tuned on the development set. The predicted probabilities are computed using the Hastie and Tibshirani's pairwise coupling method (Hastie and Tibshirani, 1998).

(iii) Classification with a non-linear discriminative classifier: Finally, we test a discriminative classifier with non-linear boundaries. We expect better results with this classifier due to a higher degree of freedom in modeling the class boundaries. On comparing results to the previous classifier, we get a sense of deviation of the non-linear class boundaries from the SVM based linear boundaries. We chose a Deep Neural Network (DNN) (Hinton et al., 2012) with sigmoidal activation as our non-linear classifier. DNNs have been used in several pattern recognition tasks such as speech recognition (Seide et al., 2011; Dahl et al., 2012) and have provided state of the art results. We train a DNN with two hidden layers and the output layer consists of three nodes with sigmoidal activation. Each output node emits a probability value for one of the three classes; laughter, filler and garbage. We perform pre-training (Dahl et al., 2012) before determining the DNN weights. The number of hidden layers and number of neurons in each hidden layer was tuned on the development set.

Results and discussion: We list the results using the three models in Table 4. The "chance" Area Under the Curve (AUC) is determined based on random assignment of 0 and 1 probability values per frame for each event.

We observe that the acoustic features considered carry distinct information about the non-verbal vocal event which distinguish them from the rest of the speech. Decoding the development and test sets using the HMM framework often outputs the garbage label for laughter and filler frames; as a large portion of the training set consists of frames with garbage label. Although, a higher weight to the events in language model leads to a better output for laughter and filler frames, but this comes at the expense of a higher false alarm rate. Between the discriminative models, we obtain better results using a non-linear boundary as compared to linear SVM boundaries given a higher degree of freedom in modeling the class distributions. The gain in the case of detecting laughters is higher as compared to that for fillers. This indicates relatively less deviation of the non-linear boundary from the SVM boundary in case of the fillers. However, in case of laughters the greater performance boost obtained using a DNN suggests that the true boundary may not be well approximated by a hyper-plane. We also observe that our classifiers obtain a higher AUC in case of fillers. This indicates that given the context independent features, fillers are comparatively more distinguishable than laughters. This happens due to inherent differences in the acoustic structure of the two events. Fillers are contiguous sounds, (e.g. um, em, eh) and laughters typically involve bursts of sounds with silence in between. This heterogeneous acoustic property of laughter makes inference more difficult. For instance, assigning a silence frame to belong to a laughter event is difficult in the absence of any context (Fig. 2). We expect better detection after incorporating contextual features as presented in the next section.

4.1.2. Frame-wise detection with contextual features

Non-verbal events such as laughters and fillers occur contiguously over long segments spanning multiple short term analysis frames. Hence the inclusion of neighboring context from surrounding frames may assist prediction as proximal acoustic properties may help resolve conflicts (e.g., in the case of silence frames within laughter). We extend the previous best system, i.e., the DNN classifier and make modifications to include feature context. We test

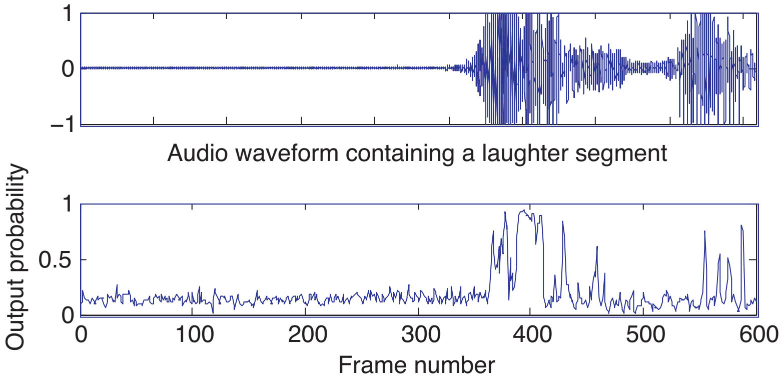

Fig. 2. An audio segment containing laughter segment at the end and the corresponding output laughter probabilities. The laughter segment in this clip (as annotated) occurs in two bursts. Notice the low laughter probability assigned to the silence frames in between the two bursts.

two methods to incorporate context from features from the neighboring frames: (i) by feature concatenation over a window followed by dimensionality reduction, and (ii) appending statistical functionals calculated over a window to the frame-wise features. We discuss each of these below.

(i) Window-wise feature concatenation and dimensionality reduction: In order to make a decision for the *n*th frame, we consider a window extending to M_x frames before and after the *n*th frame (window length: $2M_x + 1$). We concatenate features from all the frames over the window, leading to $(2M_x + 1) \times 17$ features. We determine the outcome for the *n*th frame based on these values. An increase in the number of features leads to data sparsity in the feature space. Therefore, we perform dimensionality reduction before training our classifier. We use Principal Component Analysis (PCA) (Jolliffe, 2005) to retain the maximum variance after linearly projecting the $17 \times (2M_x + 1)$ features. We train a DNN to obtain target event probabilities based on these projected features. We again downsample the garbage class before training and tune M_x, the number of principal components used and the DNN parameters on the development set. We show the classification schematic in Fig. 3. f_{FC} indicates the classifier trained with feature context.

(ii) Appending window-wise feature statistics: This scheme relies of appending the feature from the current frame with statistical functionals of features over a longer temporal window in its neighborhood. The set of features along with a linear SVM classifier was used as a baseline during the Interspeech challenge (Schuller et al., 2013). In this scheme we incorporate context by appending velocity (Δ) and acceleration (Δ^2) values calculated over the $2M_x + 1$ frame-long window to the features in the current frame. This leads to 17×3 feature values (feature + $\Delta + \Delta^2$). The practice of incorporating these contextual features (Δ and Δ^2) is widely used in applications such as speech recognition (Kingsbury et al., 1998) and language identification (Torres-Carrasquillo et al., 2002). We further calculate means and

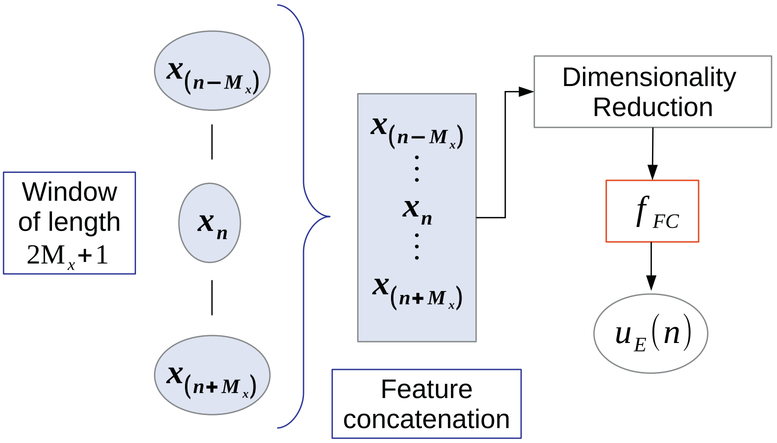

Fig. 3. Architecture for frame-wise classification based on incorporating feature context and dimensionality reduction. Feature values over a window are concatenated and projected on a lower dimensional space using PCA. f_{FC} represents the discriminative classifier trained with feature context.

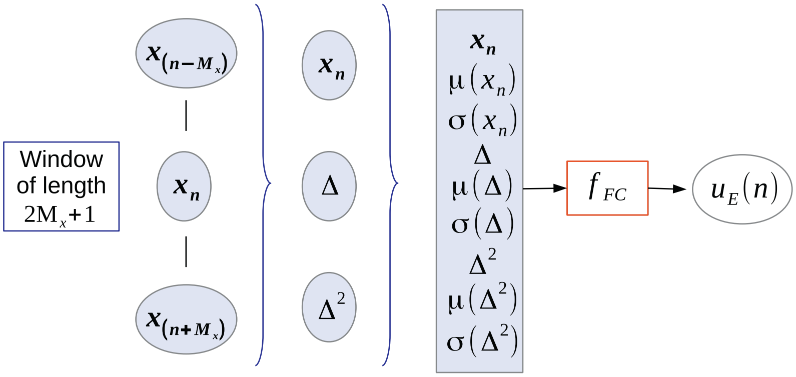

Fig. 4. Architecture for frame-wise classification based on incorporating context from feature statistics.

variances over these 51 features to obtain our final feature set. Further addition of statistical functionals (apart from Δ and Δ^2) provides additional temporal characterization and is inspired from various other previous works (Schuller et al., 2010, 2012). We train a DNN on this set of features and also report the challenge baseline results obtained using the SVM classifier for comparison. We tune M_x and the DNN parameters on the development set. Fig. 4 outlines the adopted classification scheme.

Results and discussion: We list the results using the above two classification architectures in Table 5. We also list the result from the previous best context independent classifier and the Interspeech challenge baseline results (Schuller et al., 2013) for comparison.

From the results, we observe that we obtain higher AUC values for both the events, thus validating that context helps in improving prediction. We obtain higher AUCs using the statistical functionals compared to feature concatenation. This may be due to the fact that the dimensionality reduction leads to loss of information. Also the approach with feature statistical functionals retains the features pertaining to the frame at hand. These features are otherwise projected onto a lower dimensional space in the approach involving dimensionality reduction. Given a better performance using the feature statistical functionals, we proceed with this scheme for further system development.

In this section, we explored several methodologies to extract information from a set of vocal features. Overall, non-linear boundaries on frame-wise features appended with neighboring frame information provide us the best results. Fig. 5 shows the outputs from two separate audio clips, each containing laughter and filler events. From the plots we observe that these files still contain several ‘garbage’ frames with high laughter/filler probabilities. In spite of incorporating contextual features, the estimated probabilities do not evolve smoothly. Therefore, there is potential to further improve our results after including context in the output probabilities. We address this possibility in the next session, where we account for context in the decisions obtained from the current system.

Table 5
AUC for classification using contextual features obtained by appending window-wise feature statistics and dimensionality reduction.

System used	AUC (in %)			
	Development set		Testing set	
	Laughter	Filler	Laughter	Filler
Context independent DNN	83.7	83.8	80.9	82.3
Appending features + dimensionality reduction (DNN)	86.0	88.5	84.3	84.9
Window-wise feature statistics + DNN	88.6	91.9	86.4	86.5
Window-wise feature statistics + SVM Interspeech challenge baseline (Schuller et al., 2013)	86.2	89.0	82.9	83.6

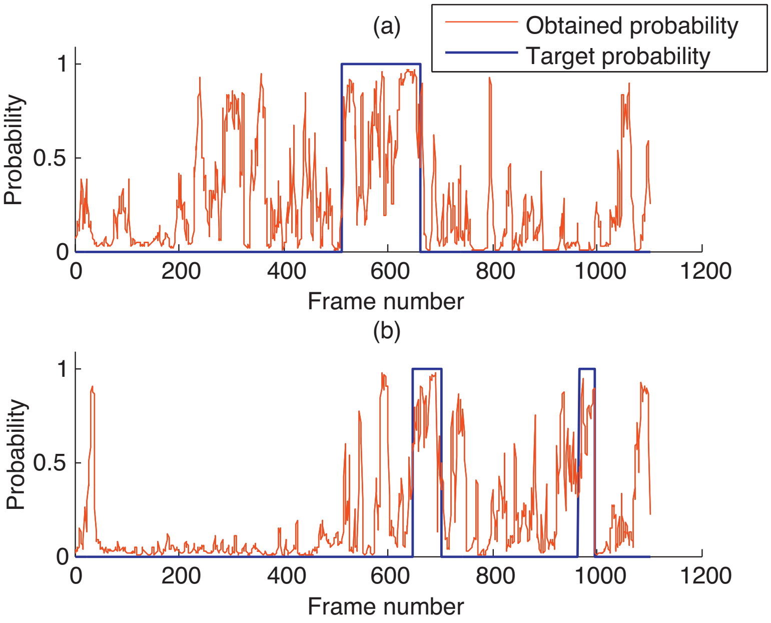

Fig. 5. Estimated and target probability values for (a) laughters and (b) fillers, using contextual features on sample test files.

4.2. *Incorporating context in probabilistic frame-wise outputs*

We propose methods to improve the previous model by incorporating context to the sequence of frame-wise probabilities (block (b) in Fig. 1). We concatenate the output frame-wise probabilities $u_E(n)$ $(n = 1 \ldots N)$ from the previous classification into a time series $U_E = \{u_E(1), \ldots, u_E(N)\}$. We perform a "smoothing" operation on U_E consisting of two steps: (i) Linear filtering, and (ii) Probability-encoding. These steps assist us in incorporating neighboring frame context decisions as discussed below.

4.2.1. *Linear filtering*

We observe that the outcomes from the above systems tend to be noisy, consisting of sharp rises and falls. We therefore design a low pass FIR filter to reduce the spikes in U, as it is unlikely that these events last only for a few frames. We determine the filtered probability $v_E(n)$ at the nth frame for an event E as shown in (1). We use a window of length $2M_u + 1$ centered at n and a_{m_u} is the filter coefficient applied to frame output at a distance of m_u from the current frame. We determine the filter coefficients using two approaches: (i) a moving average filter and (ii) a FIR filter with coefficients determined using the Minimum Mean-Squared Error (MMSE) criteria. We explain these two approaches and list the results below.

$$v_E(n) = \sum_{m_u=-M_u}^{M_u} a_{m_u} \times u_E(n + m_u) \tag{1}$$

(i) Moving Average (MA) filter: A moving average filter assigns equal values to all the coefficients, as shown in (2). For each frame, this scheme provides equal importance to the frame and its neighbors. We tune the window length parameter M_u on the development set.

$$a_{m_u} = \frac{1}{2M_u + 1}, \qquad m_u = -M_u, \ldots, M_u \tag{2}$$

(ii) Minimum Mean Squared Error (MMSE) based filter: In this filter design scheme, we find the optimal set of filter co-efficients after minimizing the Mean Squared Error (MSE, Eq. (3)) between the desired probability values and probability values obtained after filtering. MSE is a convex function with respect to (w.r.t.) the co-efficients with

Table 6
AUC after temporally filtering the time series U_E using MA and MMSE based filters. MMSE provides the best AUC, slightly better than MA filter.

System used	AUC (in %)			
	Development set		Testing set	
	Laughter	Filler	Laughter	Filler
DNN using feature statistical functionals	88.6	91.9	86.4	86.5
MA filter	97.0	95.5	92.5	89.8
MMSE filter	97.3	95.5	94.2	89.9

a global minima. Each a_{m_u} is obtained analytically by setting derivative of MSE w.r.t. a_{m_u} to zero (Eq. (4)). We tune M_u for the MMSE based filter on the development set.

$$\text{MSE} = \frac{\sum_{n \in \text{Training set}}\left(t_n - \left(\sum_{m_u=-M_u}^{M_u} a_{m_u} \times u_E(n+m_u)\right)\right)^2}{\text{Size of the training set}} \tag{3}$$

$$a_{m_u} = \underset{a_{m_u}}{argmin}(\text{MSE}) \qquad \text{at :} \qquad \frac{\partial \text{MSE}}{\partial a_{m_u}} = 0 \tag{4}$$

Results and discussion: We present the results in Table 6. We list the previous best results using contextual features for comparison.

We observe that there is a similar increase in the AUCs using the two filtering schemes.

We plot the filter coefficients in Fig. 6 and the frequency response (FFT based, Eq. (5)) of the filters in Fig. 7. A_{p_u} represents the discrete Fourier transform at the index p_u. Although the coefficients for MA and MMSE filters are different, the similarity in performance of the two filters can be explained by the similarity in their frequency response. Both these filter attenuate high frequency components in the time series U. However the MMSE filter has a slightly higher cut-off frequency and admits more high frequency components when compared to the MA filter. M_u was 50

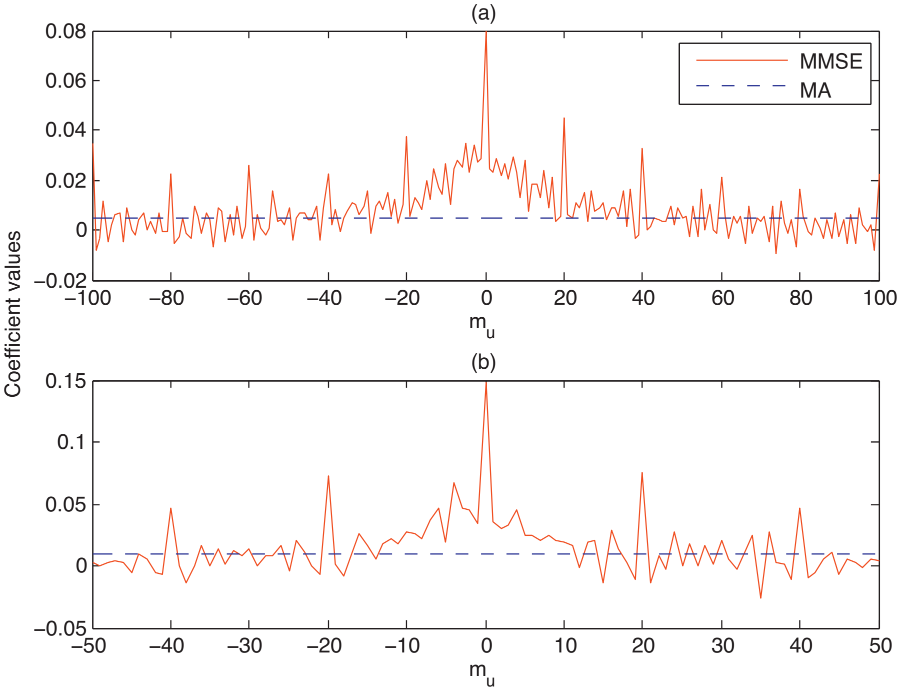

Fig. 6. Filter coefficients for the linear filters operating over the time series U_E for (a) Laughters; (b) Fillers.

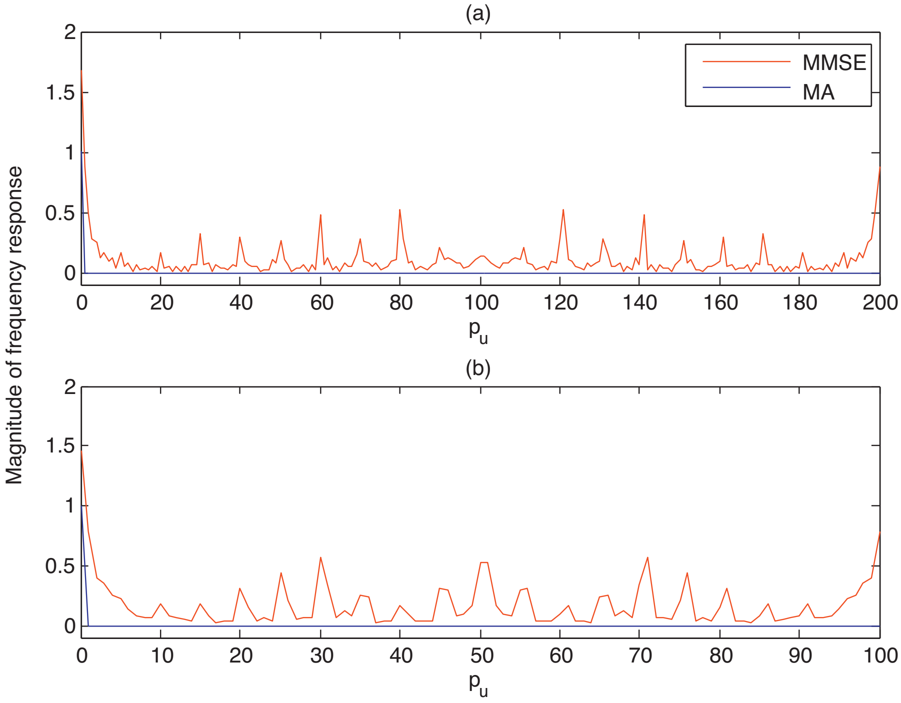

Fig. 7. Frequency response for the linear filters operating over the time series U_E for (a) Laughters; (b) Fillers.

for fillers and 100 for laughters. This suggests that the context lasts longer for laughters as compared to fillers. It may follow from the fact that mean laughter length is greater than mean filler length as is observed in Table 1.

$$A_{p_u} = \sum_{m_u=-M_u}^{M_u} a_{m_u} \exp\left(-i2\pi p_u \frac{m_u + M_u}{2M_u + 1}\right); \qquad p_u = 0\ldots, 2M_u + 1 \tag{5}$$

In the next section, we describe the probability encoding scheme.

4.2.2. Probability encoding

After processing the data through the above scheme, we pass the outputs $v_E(n)$ through an autoencoder (Baldi, 2012). The goal here is to capture any non-linear contextual dependency which the linear filters fail to capture. We define a new time series $V_E = \{v_E(1), ..., v_E(n), ..., v_E(N)\}$ consisting of the filtered outputs. The auto-encoder f_{enc} is a feed-forward neural network trained to reconstruct the target values from V_E. We use an autoencoder with a single hidden layer and sigmoidal activation on the output node. This operation reconstructs a window of inputs $\{v_E(n - M_V), \ldots, v_E(n), \ldots, v_E(n + M_V)\}$ to produce an output of the same length. The parameter M_V and number of neurons in the hidden layer are tuned on the development set. Autoencoder accounts for any non-linear dependence and is a mapping to multiple target output values (unlike linear filtering). We train a neural network encoder on a window of length $2M_v + 1$ centered at $v_E(n)$ to obtain the predictions on the same window as shown in (6).

$$\begin{aligned} \{w_E(n - M_V), &\quad \ldots, w_E(n), \ldots, w_E(n + M_V)\} \\ &= \quad f_{\text{enc}}(v_E(n - M_V), \ldots, v_E(n), \ldots, v_E(n + M_V)) \end{aligned} \tag{6}$$

Results and discussion: We list the results for auto-encoding in Table 7. We train the auto-encoder on the outputs of the MMSE filter.

Auto-encoding leads to different degrees of improvements for capturing the two events. Also, the performance is inconsistent across the data splits. On the development set, we obtain a greater increase for fillers and the pattern reverses on the test set. This is indicative of some degree of mismatch between the development and test splits. Very high AUC values on the development set indicate performance saturation and data distribution similarity between the

Table 7
AUC using the MMSE based filter and probability encoding.

System used	AUC (in %)			
	Development set		Testing set	
	Laughter	Filler	Laughter	Filler
MMSE filter	97.3	95.5	94.2	89.9
Probability encoding	97.6	96.1	95.3	90.2

development and the training set. A 1% absolute increase in AUC for laughters suggests that these events benefit more from the non-linear encoding as compared to fillers.

Overall, we observe that incorporating context from output decisions leads to a greater detection improvement in the case of laughters than fillers. This shows the importance of context, particularly in case of events with heterogeneous spatio-temporal characteristics. Fig. 8 shows the smoothed probabilities for the same set of files as shown previously in Fig. 5. We see that even though false alarms still exist for the sample files, we obtain near perfect detection in case of true positives. Also spurious isolated false alarms are largely non-existent. The false alarms are still a concern and mainly arise from similar acoustic properties between certain verbal sounds and the non-verbal events (e.g. sound “um” in umpire is similar to the filler “um”). We address this problem partially in the next section by using a ‘masking’ technique.

4.3. Masking highly probable/improbable events

As the final step, we make use of inherent properties of the probability time series to further improve our results (block (c) in Fig. 1). We develop the masking scheme based on two heuristics: (i) existence of low event probability values for extended period of time implies the absence of any event, and (ii) similarly, contiguous high event probability values implies presence of an event. We implement this strategy by developing binary masks as described below.

(i) Zeros-mask design: If there is contiguous existence of probability values below a threshold T_0 for at least a set number of K_0 frames, we mask all such probabilities by zero.

(ii) Ones-mask design: Similarly, if probability values are contiguously over a threshold T_1 for at least K_1 frames, we mask all such probabilities by one.

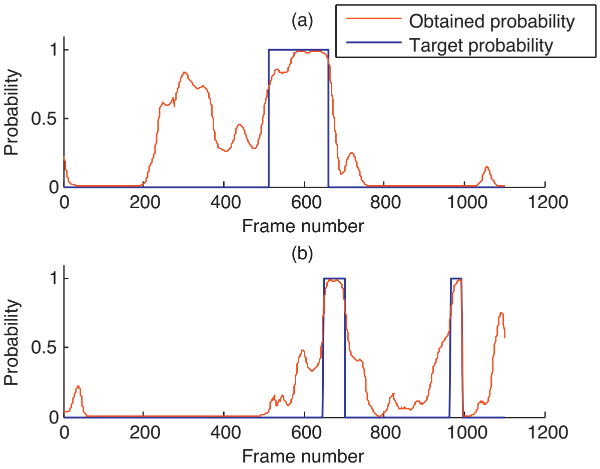

Fig. 8. Obtained and target probability values for (a) laughters and (b) fillers, after probability smoothing on sample test files.

Table 8
Detection results before and after masking the probability time series.

System used	AUC (in %)			
	Development set		Testing set	
	Laughter	Filler	Laughter	Filler
Probability encoding	97.6	96.1	95.3	90.2
Masking	97.6	96.3	95.3	90.4

The overall operation of implementing the zeros and ones masking operation is shown in (7). We tune T_0, T_1 and K_0, K_1 on the development set.

$$y_E(n) = \begin{cases} 0 & \text{if } \exists\, n \text{ such that: } w_E(n) < T_0 \, \forall\, n \in n, n+1, \ldots, n+K_0 \\ 1 & \text{if } \exists\, n \text{ such that: } w_E(n) > T_1 \, \forall\, n \in n, n+1, \ldots, n+K_1 \\ w_E(n) & \text{otherwise} \end{cases} \quad (7)$$

Results and discussion: We show the results before and after masking in Table 8.

We observe a slight increase in AUC for fillers and none for laughters. In the case of fillers, we obtain $T_0 = 0.02$ and $T_1 = 0.98$. The performance for laughters saturated in the previous step and any $T_0 > 0$ and $T_1 < 1$ led to a reduction in AUC. These threshold values suggest that the previous step involving smoothing accounted for most of the information during detection, leading to marginal gains when using additional masking. We plot the output probabilities for the chosen sample files in Fig. 9. The only visible impact is for the file containing fillers where event probabilities between frames 20 and 500 are set to zero. We perform an analysis of the results obtained after masking, and each previous operation, in the next section.

5. Analysis of features

In the experiments so far, all features are used together to make an event prediction. This makes it difficult to evaluate the contribution of individual features toward the final laughter and filler prediction outcomes. In this section, we perform two sets of experiments to address this issue. The goal of these experiments is to understand the relation

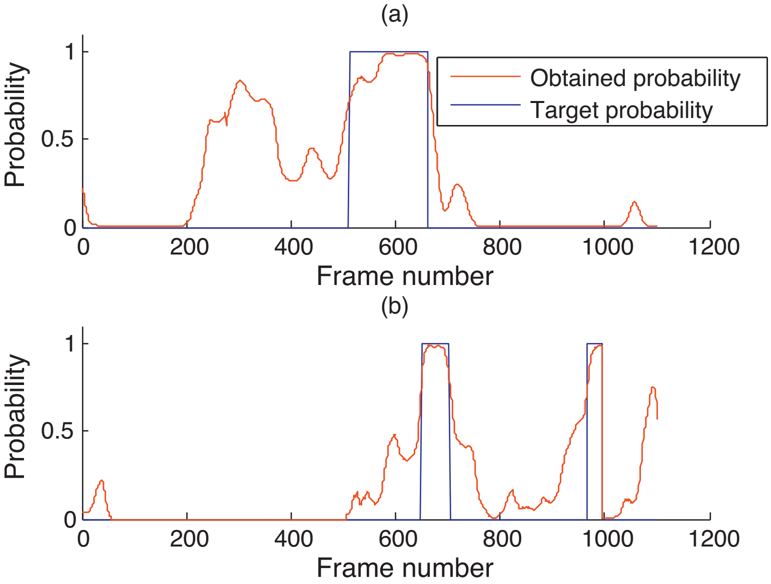

Fig. 9. Obtained and target probability values for (a) laughters and (b) fillers, after masking.

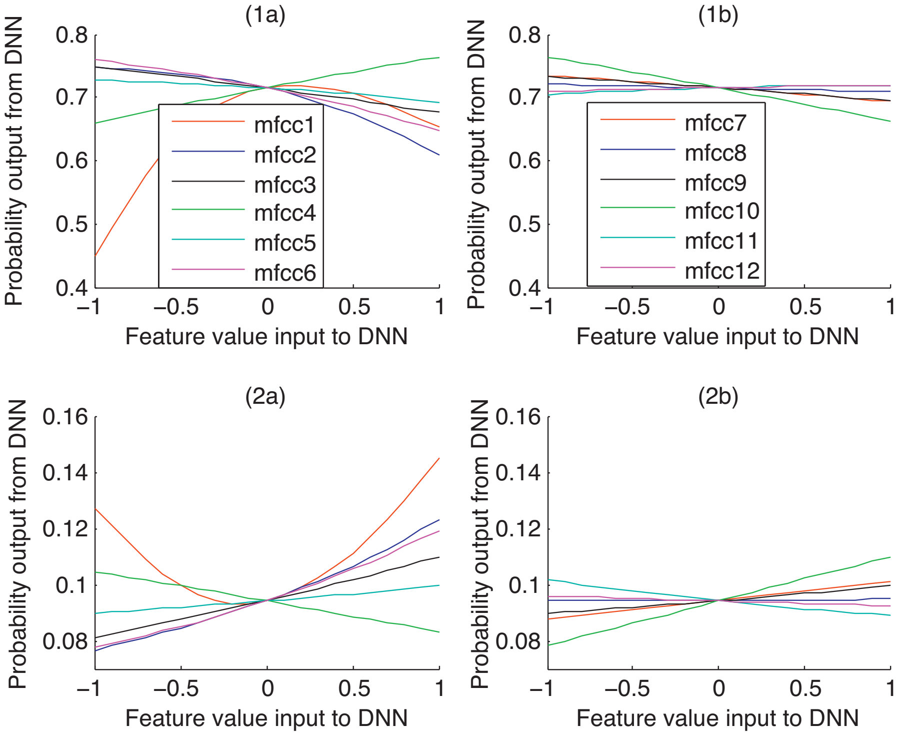

Fig. 10. Plot of event probability output by the DNN on varying the z-normalized MFCCs over one standard deviation ($\sigma = 1$). Note that while varying a single feature, all other features are set to 0. Plots 1a/2a: Plot over MFCC 1-6 for laughters/fillers. Plots 1b/2b: Plot over MFCC 7-12 for laughters/fillers.

between a feature's discriminative power and its impact on the final outcome. In the first experiment, we look at the sensitivity of the final outcome to perturbations in each single feature followed by computing the AUCs using a single feature at a time. Finally, we investigate the outcomes of these two experiments and find a significant correlation between them with further implications towards the system design.

5.1. Sensitivity analysis

In this experiment, we study the effect of each feature on the final probability outcome. During frame-wise prediction, the neurons in the DNN are activated by all feature inputs from an audio frame. However, all the features may not have similar activation patterns. Through the sensitivity analysis, we investigate the DNN activation using one feature at a time. We expect differences in activation patterns from each feature which may correlate with the feature discriminability (investigated in the next section). We perform separate analysis for the prosodic and the spectral features using our baseline DNN model trained on the 17 z-normalized features. We activate the DNN trained in Section 4.1.1(iii) using one feature at a time, while keeping all other features to be 0. We vary the chosen feature one standard deviation (−1 to +1, as it is z-normalized) around the mean (0 for z-normalized feature) and observe the probability outcome from the DNN. We plot the output probabilities for each of the features in Figs. 10 and 11. We discuss our results separately for prosodic features and MFCC coefficients below.

5.1.1. MFCC features

We show the results for the MFCC features in Fig. 10. We split the results for the 12 MFCCs in two figures for readability. From the figures, we observe that the first few MFCC co-efficients show the most dynamic range in output probability and variation is low in the second half of coefficients. The first MFCC coefficient shows the most dynamic range for both laughters and fillers indicating high sensitivity. For all coefficients except the first, a monotonic change in a value either favors laughters or fillers. The output probability trends are opposite for the two events, wherein a

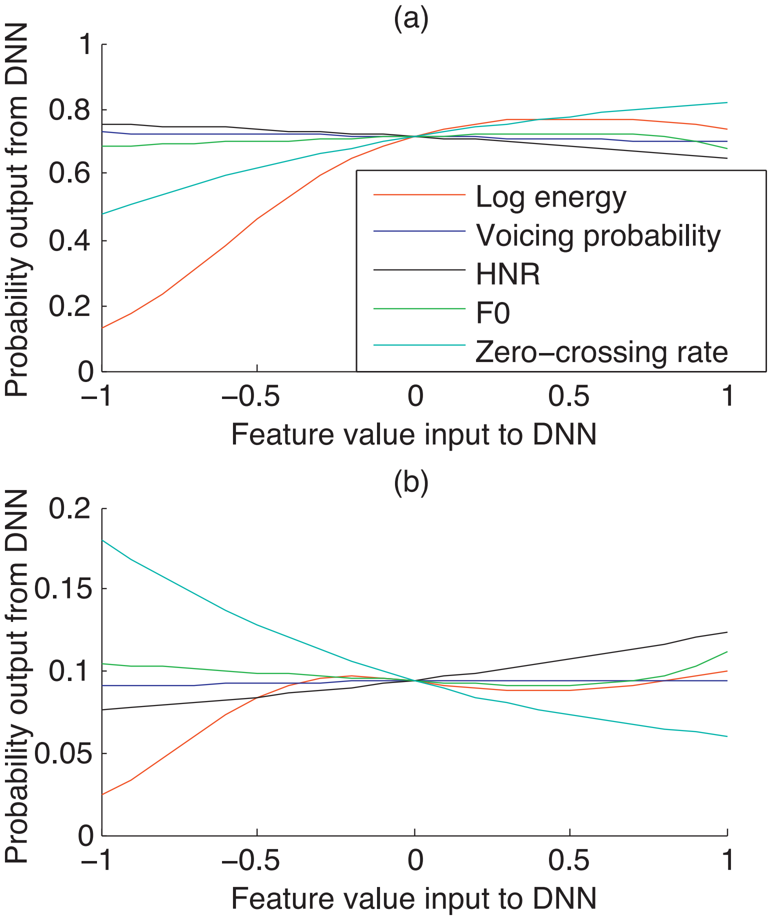

Fig. 11. Plot of assigned probabilities on varying the z-normalized prosodic features over one standard deviation. (a) and (b) Plot over prosodic features for laughters/fillers.

positive slope for laughters corresponds to a negative slope for fillers. This indicates a monotonic increase/decrease in feature values favor one event while reducing the probability of other. Note that all the curves intersect at 0 as this corresponds to a value where the neural network in not activated at all.

5.1.2. Prosodic features

The plot of assigned outputs versus variation in prosodic features is shown in Fig. 11. In case of the prosodic features, we observe the most variation in the case of log energy and zero-crossing rate. Apart from log energy, we observe the opposite trends in the slope of the curves for fillers and laughters. We observe that the curves corresponding to log energy are not monotonic, indicating a more complicated boundary for this feature. We observe patterns such as the probability of outcome increases with higher intensity (more sharply for laughters than fillers) suggesting a louder voice implies a higher laughter probability. Similar comments can be made by observing the outcome patterns with increase/decrease in a prosodic feature.

From the output patterns of the features, we observe that the trained DNN has different activation patterns for each feature. This implies that variation in features have disproportionate impact on the final probability prediction. Moreover, for several features, the output probability trends are opposite for filler and laughter events. This is expected in a discriminative model as increase in probability output for given feature values should translate to lower probability for the other. In the next experiment, we use the feature values from the dataset and predict the event labels.

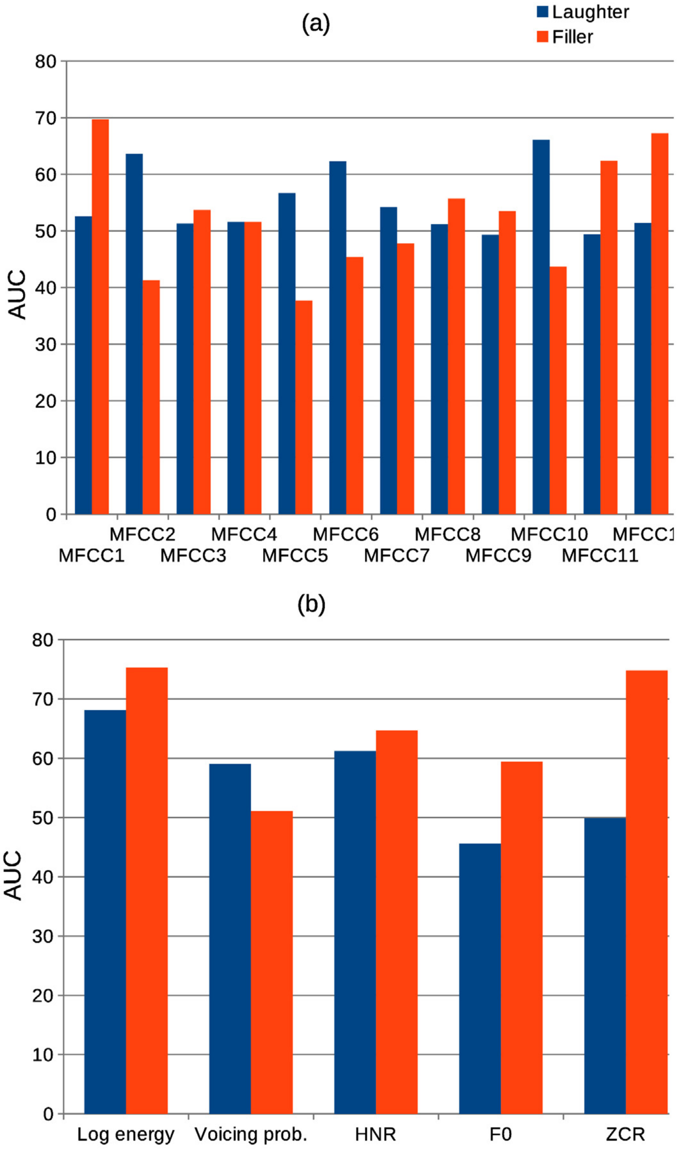

Fig. 12. AUC (in %) obtained based on a single features at a time. (a) Plot for MFCCs; (b) Plot for prosodic features.

5.2. Feature performance analysis

We analyze the performance of each feature on the test set using the same baseline DNN on 17 features (Note that this is unlike previous experiment where outputs were obtained by synthetically varying feature values). In order to predict the event probabilities, we use the values from a single feature, while setting all other features to zero. We show the corresponding AUCs for laughters and fillers in Fig. 12.

From the figures, we observe a difference in AUC patterns across the MFCC features for the two events. This shows that each frequency band carries a different amount of discriminative information for each event type. MFCC-2,6,10 perform the best for laughters and MFCC-1,11,12 for fillers. In case of the prosodic features, log energy provides the highest AUCs for both the events. Prosodic features offer a higher degree of discriminability, particularly for filler as the best two prosodic features achieve an AUC greater that 70% by themselves. We speculate that the results from

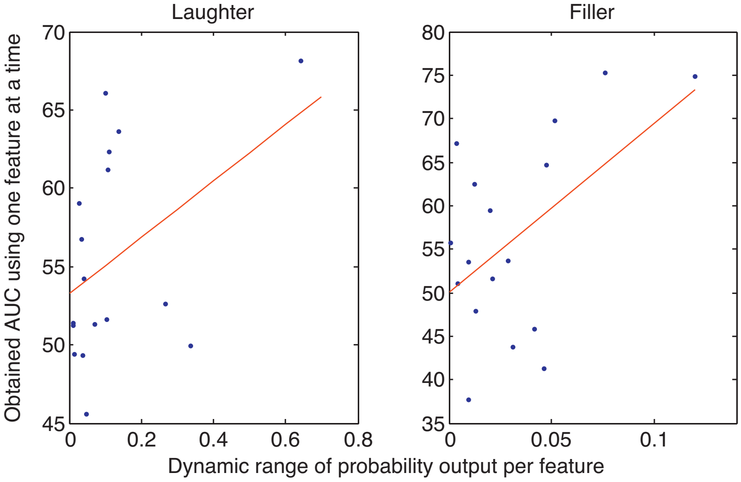

Fig. 13. Plot representing AUC and dynamic range values obtained per feature. Line in red represent the best fit using linear regression. The 17 datapoints correspond to each feature. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sensitivity analysis and feature performance analysis are related as the DNN model should tune more to better features. In the next section, we investigate the relationship between the two.

5.3. Relation between feature performance and output probability dynamic range

The relation between activation patterns in neural networks and their performance have been a subject of study in various other tasks such as face and object recognition (Lawrence et al., 1997; Le, 2013). We hypothesize that the DNN model is more sensitive to features which offer a higher discriminatory power. We test our hypothesis by performing regression analysis (Draper et al., 1966) between the dynamic range of outputs as obtained in Section 5.1 and the AUC values obtained in Section 5.2. The dynamic range obtained in sensitivity analysis is used as a proxy for the degree of activation. We fit a linear model (Eq. (8)) with dynamic range as the predictor variable for AUC values. The linear regression analysis helps us understand the general trend in between the two variables (although the relation between the two variables may not be linear). The outcome of linear fitting is shown in Fig. 13. Each of the 17 datapoints in the figure corresponds to a feature, with the x-axis representing the output dynamic range obtained during sensitivity analysis of the feature and y-axis the AUC using that feature only. Table 9 shows the statistics on the relation between the two variables. ρ represents the correlation coefficient between the AUC and the dynamic range as computed over the seventeen features.

$$\text{AUC} = \alpha_0 \times \text{Dynamic range} + \alpha_1 \tag{8}$$

The F-statistic shows that the regressor is significant at 5% level in predicting the AUC in case of fillers and at 10% level for laughters. This provides evidence of relation between a feature's discriminative power and output sensitivity to the feature. This also shows that the DNN model is more sensitve to more discriminative features. In the future, we

Table 9
Statistics of linear model for predicting AUC with dynamic range of the output probability as the regressor.

Event	Statistics w.r.t. the regressor				ρ
	α_0	Standard error	F-statistic vs constant model	p-value	
Laughter	18.1	9.64	3.51	0.08	0.44
Filler	194.4	81.9	5.63	0.03	0.53

can use this observation during DNN training by discarding nodes corresponding to features with low sensitivity. This observation may particularly be useful in case of increased dimensionality of features introduced while including the contextual features (as in Section 4.1.2).

6. Conclusion

Non-verbal vocalizations are inherent constituents of natural conversations and their interpretation can inform understanding of the communication process. We present a system for robust detection of two non-verbal events, laughter and filler, with a goal of gaining insights into the role they play during interpersonal interaction. This specific task was originally proposed as a part of the Interspeech 2013 challenge event where the initial ideas of the present study were presented (Gupta et al., 2013). Our system sequentially employs multiple signal processing methods, primarily accounting for context during the frame-wise detection as well as utilizing some inherent properties of the signal at hand. We establish an acoustic feature based context independent baseline, followed by the introduction of contextual features. Next, we incorporate context in the output decisions themselves using a "smoothing" technique. We use a linear filter and an auto-encoder for that purpose and observe performance increments. We note that we still obtain several false alarms and partially address this problem using "masking" techniques that reduce false alarms at a very low operating threshold. We observe that the performance of our system increases with each subsequent step with the contextual "smoothing" providing the most gain.

We further perform a sensitivity and performance analysis on each feature. We observe that the constituent features offer varying degree of sensitivity to perturbation and stand alone performances. It is interesting to observe that several of the sensitivity patterns follow an opposite trend when compared across the two non-verbal target events considered in this work. Also, the stand alone performance provides us a sense of which features carry the most amount of information during inference. We establish a relation between feature sensitivity and stand alone performance. We observe a positive correlation between the two and conclude that the output of the DNN model is more sensitive to more discriminative features.

Our detection scheme serves as a first step toward a robust analysis of non-verbal behaviors in vocal communication. Non-verbal vocalizations are often associated with several aspects of human behavior such as depression (Ellgring, 2007; Geerts and Bouhuys, 1998), emotions (Salovey and Mayer, 1989; Fabri et al., 1999) as well as a marker of the overall quality of interaction (Burgoon et al., 2002; Gabbott and Hogg, 2000). Detection of non-verbal events can facilitate further investigations such as studying their patterning with respect to specific communication settings and goals. This can be particularly beneficial in diagnostic settings related with depression, autism and other such disorders (Geerts and Bouhuys, 1998; Ulrich and Harms, 1985; Gotham et al., 2007). Furthermore, our current system can be further expanded to incorporate other non-verbal event types. The proposed framework is general and event-specific alterations can be introduced. Even though our system successfully infer a majority of 'garbage' frames as not belonging to an event, false alarms pose challenges given a lower frequency of occurrence of the target events. Additional measures can be introduced to address this issue and one suggested approach is to create a complementary modeling of the 'garbage' frame themselves, which are composed of speech and silence. A long term goal may also include characterizing the distribution of these events with adapting the output probabilities of our current system based on the distribution prior.

Acknowledgment

This work was done as part of the Defense Advanced Research Projects Agency Broad Operational Language Translation (DARPA-BOLT) project. The authors would like to thank DARPA for their support.

References

An, G., Brizan, D.-G., Rosenberg, A., 2013. Detecting laughter and filled pauses using syllable-based features. In: INTERSPEECH, pp. 178–181.

Argyle, M., Salter, V., Nicholson, H., Williams, M., Burgess, P., 1970. The communication of inferior and superior attitudes by verbal and non-verbal signals. Br. J. Soc. Clin. Psychol. 9 (3), 222–231.

Argyle, M., 1972. Non-verbal communication in human social interaction. In: Hinde, R.A. (Ed.), Non-verbal Communication. Cambridge University Press, Oxford, England, xiii, 443 pp.

Bachorowski, J.-A., Smoski, M.J., Owren, M.J., 2001. The acoustic features of human laughter. J. Acoust. Soc. Am. 110 (3), 1581–1597.

Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures. In: ICML Unsupervised and Transfer Learning, pp. 37–50.

Bavelas, J.B., Chovil, N., 2006. Nonverbal and verbal communication: hand gestures and facial displays as part of language use in face-to-face dialogue. In: The Sage Handbook of Nonverbal Communication., pp. 97–115.

Boland, J.P., Follingstad, D.R., 1987. The relationship between communication and marital satisfaction: a review. J. Sex Marital Ther. 13 (4), 286–313.

Bone, D., Lee, C.-C., Chaspari, T., Black, M.P., Williams, M.E., Lee, S., Levitt, P., Narayanan, S., 2013. Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand. In: INTERSPEECH.

Bone, D., Lee, C.-C., Black, M.P., Williams, M.E., Lee, S., Levitt, P., Narayanan, S., 2014. The psychologist as an interlocutor in autism spectrum disorder assessment: insights from a study of spontaneous prosody. J. Speech Lang. Hear. Res. 57 (4), 1162–1177.

Bowers, D., Bauer, R.M., Heilman, K.M., 1993. The nonverbal affect lexicon: Theoretical perspectives from neuropsychological studies of affect perception. Neuropsychology 7 (4), 433.

Brueckner, R., Schuller, B., 2013. Hierarchical neural networks and enhanced class posteriors for social signal classification. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 362–367.

Brueckner, R., Schulter, B., 2014. Social signal classification using deep blstm recurrent neural networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4823–4827.

Burgoon, J.K., Bonito, J.A., Ramirez, A., Dunbar, N.E., Kam, K., Fischer, J., 2002. Testing the interactivity principle: Effects of mediation propinquity and verbal and nonverbal modalities in interpersonal interaction. J. Commun. 52 (3), 657–677.

Cai, R., Lu, L., Zhang, H.-J., Cai, L.-H.,2003. Highlight sound effects detection in audio stream. In: Proceedings of 2003 International Conference on Multimedia and Expo, ICME'03, Vol. 3. IEEE, III-37.

Candea, M., Vasilescu, I., Adda-Decker, M., et al., 2005. Inter-and intra-language acoustic analysis of autonomous fillers. In: Proceedings of DISS 05. Disfluency in Spontaneous Speech Workshop, pp. 47–52.

Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20 (3), 273–297.

Cunningham, M.R., 1977. Personality and the structure of the nonverbal communication of emotion. J. Personal. 45 (4), 564–584.

Curcio, F., 1978. Sensorimotor functioning and communication in mute autistic children. J. Autism Child. Schizophr. 8 (3), 281–292.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech Lang. Process. 20 (1), 30–42.

Draper, N.R., Smith, H., Pownell, E., 1966. Applied Regression Analysis, Vol. 3. Wiley, New York.

Ellgring, H., 2007. Non-Verbal Communication in Depression. Cambridge University Press.

Eyben, F., Wöllmer, M., Schuller, B.,2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the International Conference on Multimedia. ACM, pp. 1459–1462.

Fabri, M., Moore, D.J., Hobbs, D.J.,1999. The emotional avatar: non-verbal communication between inhabitants of collaborative virtual environments. In: Gesture-Based Communication in Human-Computer Interaction. Springer, pp. 269–273.

Forney Jr., G.D., 1973. The viterbi algorithm. Proc. IEEE 61 (3), 268–278.

Funahashi, K.-i., Nakamura, Y., 1993. Approximation of dynamical systems by continuous time recurrent neural networks. Neural Netw. 6 (6), 801–806.

Furlow, F.B., 1997. Human neonatal cry quality as an honest signal of fitness. Evol. Hum. Behav. 18 (3), 175–193.

Gabbott, M., Hogg, G., 2000. An empirical investigation of the impact of non-verbal communication on service evaluation. Eur. J. Mark. 34 (3/4), 384–398.

Geerts, E., Bouhuys, N., 1998. Multi-level prediction of short-term outcome of depression: non-verbal interpersonal processes cognitions and personality traits. Psychiatry Res. 79 (1), 59–72.

Goodwin, J., Jasper, J.M., Polletta, F., 2009. Passionate Politics: Emotions and Social Movements. University of Chicago Press.

Gotham, K., Risi, S., Pickles, A., Lord, C., 2007. The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. J. Autism Dev. Disord. 37 (4), 613–627.

Gupta, R., Lee, C.-C., Bone, D., Rozga, A., Lee, S., Narayanan, S., 2012. Acoustical analysis of engagement behavior in children. In: Workshop on Child Computer Interaction.

Gupta, R., Lee, C.-C., Narayanan, S., 2012. Classification of emotional content of sighs in dyadic human interactions. In: 2012 IEEE International Conference on IEEE Acoustics, Speech and Signal Processing (ICASSP), pp. 2265–2268.

Gupta, R., Audhkhasi, K., Lee, S., Narayanan, S., 2013. Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In: Proceedings of Interspeech, Lyon, France, pp. 173–177.

Gupta, R., Georgiou, P., Atkins, D., Narayanan, S., 2014. Predicting clients inclination towards target behavior change in motivational interviewing and investigating the role of laughter. In: Proceedings of Interspeech.

Halberstadt, A.G., 1986. Family socialization of emotional expression and nonverbal communication styles and skills. J. Personal. Soc. Psychol. 51 (4), 827.

Harris, M., Jones, D., Brookes, S., Grant, J., 1986. Relations between the non-verbal context of maternal speech and rate of language development. Br. J. Dev. Psychol. 4 (3), 261–268.

Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (Eds.), Advances in Neural Information Processing Systems, Vol. 10. MIT Press.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. Signal Process. Mag. IEEE 29 (6), 82–97.

Hogan, K., Stubbs, R., 2003. Can't Get Through: Eight Barriers to Communication. Pelican Publishing.

Isbister, K., Nass, C., 2000. Consistency of personality in interactive characters: verbal cues non-verbal cues and user characteristics. Int. J. Hum. Comput. Stud. 53 (2), 251–267.

Johannesen, R.L., 1971. The emerging concept of communication as dialogue. Q. J. Speech 57 (4), 373–382.

Jolliffe, I., 2005. Principal Component Analysis. Wiley Online Library.

Kahn, M., 1970. Non-verbal communication and marital satisfaction. Fam. Process 9 (4), 449–456.

Kaya, H., Erçetin, A.M., Salah, A.A., Gürgen, F., 2013. Random forests for laughter detection. In: Proceedings of Workshop on Affective Social Speech Signals-in conjunction with the INTERSPEECH.

Kennedy, L.S., Ellis, D.P., 2004. Laughter detection in meetings. In: NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, National Institute of Standards and Technology, pp. 118–121.

Kingsbury, B.E., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Commun. 25 (1), 117–132.

Krikke, T.F., Truong, K.P., 2013. Detection of nonverbal vocalizations using Gaussian mixture models: looking for fillers and laughter in conversational speech. In: Proceedings of Interspeech, Lyon, France, pp. 163–167.

Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lanzetta, J.T., Kleck, R.E., 1970. Encoding and decoding of nonverbal affect in humans. J. Personal. Soc. Psychol. 16 (1), 12.

Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D., 1997. Face recognition: a convolutional neural-network approach. IEEE Trans. Neural Netw. 8 (1), 98–113.

Le, Q.V., 2013. Building high-level features using large scale unsupervised learning. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8595–8598.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M., 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Trans. Audio Speech Lang. Process. 14 (5), 1526–1540.

Mehrabian, A., Ferris, S.R., 1967. Inference of attitudes from nonverbal communication in two channels. J. Consult. Psychol. 31 (3), 248.

Mowrer, D.E., LaPointe, L.L., Case, J., 1987. Analysis of five acoustic correlates of laughter. J. Nonverbal Behav. 11 (3), 191–199.

Mundy, P., Sigman, M., Ungerer, J., Sherman, T., 1986. Defining the social deficits of autism: the contribution of non-verbal communication measures. J. Child Psychol. Psychiatry 27 (5), 657–669.

O'sullivan, T., Hartley, J., Saunders, D., Montgomery, M., Fiske, J., 1994. Key Concepts in Communication and Cultural Studies. Routledge, London.

Pammi, S., Chetouani, M., 2013. Detection of social speech signals using adaptation of segmental hmms. In: Workshop on Affective Social Speech Signals, Grenoble.

Piccardi, M.,2004. Background subtraction techniques: a review. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, Vol. 4. IEEE, pp. 3099–3104.

Pierre-Yves, O., 2003. The production and recognition of emotions in speech: features and algorithms. Int. J. Hum. Comput. Stud. 59 (1), 157–183.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlí ček, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The kaldi speech recognition toolkit. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2011).

Poyatos, F., 1992. Advances in Non-Verbal Communication: Sociocultural, Clinical, Esthetic and Literary Perspectives. John Benjamins Publishing Company.

Rabiner, L., Juang, B.-H., 1986. An introduction to hidden markov models. IEEE ASSP Mag. 3 (1), 4–16.

Ruesch, J., Kees, W., 1956. Nonverbal Communication: Notes on the Visual Perception of Human Relations. University of California Press.

Salamin, H., Polychroniou, A., Vinciarelli, A., 2013. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4282–4287.

Salovey, P., Mayer, J.D., 1989. Emotional intelligence. Imagin. Cogn. Personal. 9 (3), 185–211.

Schuller, B., Eyben, F., Rigoll, G., 2008. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In: Perception in Multimodal Dialogue Systems. Springer, pp. 99–110.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S.S., 2010. The interspeech 2010 paralinguistic challenge. In: INTERSPEECH, pp. 2794–2797.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Van Son, R., Weninger, F., Eyben, F., Bocklet, T., et al., 2012. The interspeech 2012 speaker trait challenge. In: INTERSPEECH.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S., 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of Interspeech.

Seide, F., Li, G., Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks. In: Interspeech, pp. 437–440.

Soltysik, S., Jelen, P., 2005. In rats, sighs correlate with relief. Physiol. Behav. 85 (5), 598–602.

Streeck, J., Knapp, M.L., 1992. The interaction of visual and verbal features in human communication. Advances in Nonverbal Communication, 3–23.

Sundaram, S., Narayanan, S., 2007. Automatic acoustic synthesis of human-like laughtera). J. Acoust. Soc. Am. 121 (1), 527–535.

Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller Jr., J.R., 2002. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: INTERSPEECH.

Truong, K.P., Van Leeuwen, D.A., 2005. Automatic detection of laughter. In: INTERSPEECH, pp. 485–488.

Tu, Z.,2005. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005, Vol. 2. IEEE, pp. 1589–1596.

Ulrich, G., Harms, K., 1985. A video analysis of the non-verbal behaviour of depressed patients before and after treatment. J. Affect. Disord. 9 (1), 63–67.

Várallyay Jr., G., Benyó, Z., Illényi, A., Farkas, Z., Kovács, L., 2004. Acoustic analysis of the infant cry: classical and new methods. In: 26th Annual International Conference of the IEEE, Vol. 1, IEEE Engineering in Medicine and Biology Society, 2004. IEMBS'04, pp. 313–316.

Vasilescu, I., Candea, M., Adda-Decker, M., et al., 2005. Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages. In: Proceedings of Interspeech.

Vettin, J., Todt, D., 2004. Laughter in conversation: features of occurrence and acoustic structure. J. Nonverbal Behav. 28 (2), 93–115.

Vlemincx, E., Taelman, J., Van Diest, I., Van den Bergh, O., 2010. Take a deep breath: the relief effect of spontaneous and instructed sighs. Physiol. Behav. 101 (1), 67–73.

Wagner, J., Lingenfelser, F., André, E., 2013. Using phonetic patterns for detecting social cues in natural conversations. In: INTERSPEECH, pp. 168–172.

Zuckerman, M., Lipets, M.S., Koivumaki, J.H., Rosenthal, R., 1975. Encoding and decoding nonverbal cues of emotion. J. Personal. Soc. Psychol. 32 (6), 1068.