AUDIOVISUAL CLASSIFICATION OF VOCAL OUTBURSTS IN HUMAN CONVERSATION USING LONG-SHORT-TERM MEMORY NETWORKS

Florian Eyben[1]*,Stavros Petridis*[2]*,Björn Schuller*[1]*,George Tzimiropoulos*[2]*,Stefanos Zafeiriou*[2]*,Maja Pantic*[2,3]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Department of Computing, Imperial College London, United Kingdom
[3]EEMCS, University of Twente, Netherlands
eyben@tum.de

ABSTRACT

We investigate classification of non-linguistic vocalisations with a novel audiovisual approach and Long Short-Term Memory (LSTM) Recurrent Neural Networks as highly successful dynamic sequence classifiers. As database of evaluation serves this year's Paralinguistic Challenge's Audiovisual Interest Corpus of human-to-human natural conversation. For video-based analysis we compare shape and appearance based features. These are fused in an early manner with typical audio descriptors. The results show significant improvements of LSTM networks over a static approach based on Support Vector Machines. More important, we can show a significant gain in performance when fusing audio and visual shape features.

Index Terms— Non-linguistic Vocalisations, Laughter, Audiovisual Processing, Long Short-Term Memory

1. INTRODUCTION

Although cognitive scientists were unable to identify a set of vocal cues that reliably discriminate among affective states and attitudes, listeners seem to be rather accurate in decoding some non-basic affective states such as distress, anxiety, boredom, and sexual interest from non-linguistic vocalisations like laughs, cries, sighs, and yawns. This finding instigated the research on automatic analysis of vocal non-linguistic expressions. More generally, these vocal episodes are highly relevant behavioral patterns for recognition of human affect, social signals, and personality traits [1] and, in turn, they play an important role to a multiplicity of applications including Automatic Speech Recognition (to avoid substitutions with in-vocabulary linguistic events), affective and socially-aware computing, and future communication and media retrieval systems. While a growing number of efforts towards automatic recognition of non-linguistic vocal outbursts is recently reported, most of these are based only on audio signals and aimed at automatic laughter recognition [2–4]. Laughter is a highly variable acoustic signal [5,6] which is accompanied by a facial expression. Therefore lately, promising successes are observed by audiovisual assessment. Since it has been shown by several experimental studies in either psychology or signal processing that integrating the information from audio and video leads to an improved performance of human behaviour recognition, few pioneering efforts towards audiovisual recognition of non-linguistic vocal outbursts have been recently reported including mainly automatic classification of audiovisual laughter episodes [7–9] but also audiovisual analysis of cries [10].

A major challenge in studying laughter and related vocal non-linguistic outbursts, especially in an audiovisual way, is the lack of data. Since laughter and its like usually occur in social situations, when people are in groups, it is not easy to obtain clear recording of individual spontaneous and natural expressions. Consequently, meeting corpora are commonly used which are different in each work and as a result direct comparability of findings is usually very limited.

Similar audio features as for speech recognition, like MFCC, PLP or RASTA-PLP, [8, 9, 11] are popular, but different features like spectral power, entropy [12], and modulation spectrum features [13] have also been considered. In terms of visual features, mostly shape features have been employed [8, 9, 11], since the use of spontaneous data from meetings with large head movements and non-frontal poses makes it difficult to apply appearance features. Different types of visual features have also been considered, like face and body movement features [13].

Many research challenges are yet to be investigated including handling continuous stream of audiovisual data to be analysed for presence of target vocal outbursts, dealing with presence of noise, (dynamic) reverberation, and package loss in transmission, and handling recordings of multiple speakers captured by a single or limited sensor(s).

In this contribution, we want to face in particular natural conversational behaviour and investigate the gain by audiovisual fusion stemming from a slightly broadened basis of audio descriptors and considering shape and appearance based video features. These are fused early and we further introduce the Long-Short-Term Memory (LSTM) paradigm for their classification under context exploitation. The targets of interest include conversational consent and hesitation—which to our knowledge have not been pursued in an audiovisual manner, yet—and laugher as opposed to 'garbage' in the sense of speech and other vocalisation as breathing or coughing.

The remainder of this paper is structured as follows: we first introduce our methodology in Sec. 2, experimental protocol and results in Sec. 3 before drawing our conclusions in Sec. 4.

2. AUDIOVISUAL VOCAL OUTBURST CLASSIFICATION

This section describes the proposed approach for multimodal classification of non-linguistic vocalisations using Long Short-Term

At the time of writing, the third author was a visiting researcher at the Imperial College London's Department of Computing. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

Fig. 1: Original frame (left) with tracked facial points and warped frontal face example (right). TUM AVIC corpus.

Memory Recurrent Neural Networks (RNN). In order to perform multimodal fusion of audio and video cues, we use the concept of early fusion, that is, feature level fusion where visual and acoustic features are concatenated for every frame. In the following we first describe the video and audio related features. Next, we give a short introduction to the concept of LSTM as used for contextual modelling on a frame-level.

2.1. Video features

In this contribution, we use both shape and appearance features. Initially, we track 20 facial points using the Patras-Pantic particle filtering tracking scheme [14]. These points are the corners of the eyebrows (2 points), the eyes (4 points), the nose (3 points), the mouth (4 points), and the chin (1 point)—for an example refer to Fig. 1. For each video segment containing K frames, we obtain a set of K vectors containing 2D coordinates of the 20 points. Employing a Point Distribution Model (PDM), by applying PCA to the matrix of these K vectors, head movement can be decoupled from facial expression. Using the approach proposed in [15], the facial expression movements are encoded by the projection of the tracking points' coordinates to the N principal components (PCs) of the PDM which correspond to facial expressions. So our shape features are the projection of the 20 points to the 6 PCs which were found to correspond to facial expressions (PCs 5 to 10), and are extracted at the video frame rate, i. e., 25 fps. Further details of the feature extraction procedure can be found in [15]. This set is referred to as '*Shape*' in the ongoing.

We further consider appearance features obtained as follows: First, we registered and cropped all faces from all subjects. More specifically, from the set of 20 tracked facial fiducial points, we selected 5 points corresponding to the 4 eye corners and the tip of the nose which remain relatively stable and invariant to facial deformations (see Fig. 1). We then employ the coordinates of these tracked points as well as the coordinates of these points in a reference coordinate system to solve for an affine transformation. We utilise the estimated transform to warp each face to the reference frame. Finally, all faces are re-sampled to dimension 64×64. Once all faces are warped, we obtain appearance features by applying PCA to image gradients [16] and keeping the first 30 dimensions. This set is referred to as '*Appear*' in the ongoing. An example is shown in Fig. 1. The actual frame, with the head rotated, can be seen on the left. On the right, the warped face, which is now frontal, is shown. There is evidence in psychological literature that changes in the upper face appearance are present particularly in laughter [17], therefore we keep the entire face region when extracting appearance features.

2.2. Audio features

We decided for a compact set of 9 acoustic low-level descriptors, which are commonly used for related tasks such as emotion recognition and speech recognition (cf. Table 1) and their respective first and second order delta regression coefficients. We chose to use only Perceptual Linear Prediction Cepstral Coefficients (PLP-CC) 1–5 instead of coefficient 1–12 as is usual for automatic speech recognition applications in order to keep the dimensionality of the acoustic feature set similar to the shape based set and as it is known that these suffice for non-linguistic assessment.

Acoustic features have been calculated using our open-source extractor openSMILE [18] at 100 fps. The full set is 27 dimensional after addition of first and second order delta regression coefficients and will be referred to as '*Audio*' in the ongoing.

Acoustic Low-level Descriptors (9)
Perceptual Linear Prediction Cepstral Coefficients (PLP-CC) 1–5
Logarithmic Energy
Loudness
Fundamental Frequency (F_0)
Probability of Voicing

Table 1: Set of 9 acoustic low-level descriptors.

2.3. Long Short-Term Memory Recurrent Neural Networks

We will only briefly describe the theory of LSTM networks and motivate their use for non-linguistic vocalisations recognition. For a detailed discussion of the LSTM concept and network architecture, we refer to [19].

In principle LSTM networks are an extension of recurrent neural networks, which in turn are standard neural networks with recurrent (feed-back) connections. However, they are equipped with an enhanced memory feature. RNNs consist of one input, one output and one or more hidden layer(s). The recurrent connections allow the network theoretically to map from the entire history of previous inputs to an output. They form a kind of memory, which allows input values to persist in the hidden layer(s) and influence the network output in the future. Although by that RNNs have access to all past information in theory, the actual range of context is limited to a few frames due to the vanishing gradient problem: The influence of an input value decays or blows up exponentially over time.

To overcome this deficiency, the LSTM concept was introduced: In an LSTM hidden layer, the non-linear units are extended to LSTM memory blocks. Each block contains one or more linear memory units, whose internal state is maintained by a recurrent connection with constant weight 1.0, enabling the unit to store information over arbitrary periods of time. The input, output, and internal state of the memory units are controlled by multiplicative gate units, which correspond to write, read, and reset operations. The gates are connected to the input layer as well as recurrently to the output layer. During network training, the weights for all connections, including the gate units, are optimised such that the network—ideally—automatically learns when to store, use, or discard information acquired from previous inputs or outputs. This makes LSTM RNN useful for many connected sequence classification tasks where context is important but exact nature of the dependencies is unknown a priori. LSTM RNN have been successfully used for a great variety of applications often outperforming more traditional sequence classifiers such as Hidden Markov Models.

3. EXPERIMENTS AND RESULTS

3.1. Data and Protocol

We prepared a data set based on the TUM Audio-Visual Interest Corpus (TUM AVIC). It consists of 3 901 turns of natural human-to-human conversational speech of a product presentation, spoken by 21 subjects (10 of them female). The total recording time for males resembles 5:14:30 h with 1 907 turns, for females 5:08:00 h with 1 994 turns, respectively. The spoken content, including non-linguistic vocalisations, is transcribed on the word level. For a detailed description of TUM AVIC we refer to [20].

We follow the official partitioning of the corpus as was used for the INTERSPEECH 2010 Paralinguistic Challenge [21]. By that, there are 718 non-linguistic vocalisations in the evaluation set, and 1 573 non-linguistic vocalisations in the training set with more than 3 frames (instances with less than 3 video frames (120 ms) were discarded to avoid processing problems). These numbers exclude the class "breath", i. e., they include the classes (instances per train/evaluation): GAR BAGE (420 / 161), CON SENT (218 / 91), HES ITATION (731 / 403), LAU GHTER (204 / 63).

In the experiments presented here we consider isolated non-linguistic vocalisations as in [22]. At being we thus do not consider detection of non-linguistic vocalisations within continuous utterances in order to properly assess the discriminative abilities of LSTM RNN and the feasibility of audiovisual fusion for the task. However, as discussed below, the LSTM RNN approach is in principle also capable of detecting events in continuous utterances through a slight modification of the network output representation.

For this task of isolated non-linguistic vocalisations classification we investigate—as stated—the performance of visual shape descriptors, visual appearance based features, and audio low-level descriptors individually as well as in all possible combinations using feature-level fusion. Audio and visual features are extracted at different frame rates, 100 fps and 25 fps, respectively, so visual features are upsampled simply by copying each feature vector 4 times in order to match the audio frame rate. We compare dynamic, frame-wise classification with LSTM RNN followed by weighted majority voting to a static classification approach where low-level descriptor contours are mapped to a fixed length vector via functionals and Support Vector Machines are employed in a subsequent classification step for reference. For all experiments the classifier of choice has been trained on the joint data from the TUM AVIC training and development set, which we will refer to as *training data* in the ongoing. Evaluations have been conducted on the TUM AVIC evaluation set.

We tested several LSTM configurations and topologies by training on the TUM AVIC training set and evaluating on the development set. We found the best configuration to have a single hidden layer with 125 LSTM memory blocks with one cell each. The networks used in this paper have an input layer with N_i linear summation input units, a hidden layer with 125 LSTM blocks with one memory cell each, and a soft-max output layer with 4 outputs. In our case we require the 4 outputs for the 4 classes we wish to detect. Due to the soft-max constraint the sum of all the outputs in the output layer always equals 1, thus the values of the 4 outputs can be regarded as probabilities that a certain frame belongs to the respective class. An alternative way is not to use a soft-max output layer, but a standard sigmoid layer. In this case we require only 3 output units, one for each class, except the GARBAGE class. GARBAGE is represented (under ideal circumstances) as all outputs being 0. To effectively discriminate between GARBAGE and one of the 3 other classes in this case, we have to apply a detection threshold to the outputs and then choose the output with the maximum value above the threshold. If no output is above this threshold, no non-linguistic vocalisation is detected. This approach is suitable for detecting non-linguistic vocalisations in continuously spoken utterances and will be investigated in follow-up work.

Functionals (7)
Extremes (maximum, minimum value)
Range (maximum – minimum value)
Arithmetic mean
Standard deviation
Skewness, Kurtosis

Table 2: Set of 7 functionals used to convert low-level feature contours of variable length to a fixed length vector for static classification with SVM.

For multimodal classification of isolated non-linguistic vocalisations, let each vocalisation be represented by a sequence $\underline{X}$ of feature vectors $\underline{x}_j$. An LSTM RNN is trained as a frame-wise classifier, i. e., a target representing the ground-truth class label l of each vocalisation $\underline{X}$ is assigned to all frames $\underline{x}_j$ belonging to this vocalisation for training of the network. During evaluation majority voting is applied to assign a single label to the sequence: The sum of each network output over the whole sequence is computed and the class label corresponding to the output with the highest sum is chosen as sequence label.

LSTM RNN are trained using resilient propagation instead of standard gradient descent. For more details on these methods, please refer to [19]. Batch learning is applied, i. e., the network weights are not updated after processing every single training instance, but only after processing the whole training data.

The static classification approach is similar to the one introduced in [22], except that we use a different feature set and the official corpus partitioning from the INTERSPEECH 2010 Paralinguistic Challenge instead of 3-fold cross validation. Moreover, for a fair comparison, the feature set used herein is based on the same low-level acoustic descriptors as used for the proposed LSTM RNN approach. Again, we consider the low-level descriptor sets *Audio*, *Shape*, *Appear* as well as all combinations of these. We then apply a small set of statistical functionals (table 2) to the (fused) set of low-level descriptors' contours.

3.2. Experimental Results

We report weighted (WAR, i. e., accuracy) and unweighted average recall (UAR) rates for all experiments in Table 3. As can be seen, LSTM networks perform better on unbalanced data with respect to UAR. The overall best result—by fusion of audio and shape features—is also obtained by LSTM on the frame level. Interestingly, the accuracy is not raised by addition of shape features, yet the unweighted accuracy is highly significantly boosted. Comparing shape and appearance features, the latter not only clearly fall behind, but their inclusion worsens results in any combination, as they appear close to chance level. A possible explanation for that is high registration errors due to out of plane rotations. This type of rotation occurs sometimes and a fully frontal face cannot be obtained by an affine transformation. Appearance features are more sensitive to this type of errors and this may lead to inferior performance than shape features. In addition, the extraction of appearance features from the entire face, which is supported by evidence in psychology as mentioned above (section 2.1), may also degrade the performance. Therefore the use of appearance features extracted only from the lower part of the face, as it is common in audiovisual

[%]	LSTM		SVM	
Features	UAR	WAR	UAR	WAR
Appear	32.5	50.0	31.8	60.0
Shape	48.4	56.1	39.6	60.2
Shape+Appear	40.8	51.8	39.2	58.2
Audio	64.6	73.5	59.1	72.4
Audio+Appear	60.3	64.2	59.4	72.1
Audio+Shape	**72.0**	**73.5**	60.5	72.4
Audio+Shape+Appear	64.3	63.1	62.7	74.2

Table 3: Results for multimodal non-lingusitic vocalisation classification on TUM AVIC. Low-level feature fusion of various sets: appearance based features (Appear), shape features (Shape), and audio features (Audio). Weighted average (WAR) and unweighted average (UAR) of class-wise recall rates. Details in the text.

[%] **as** →	GAR	CON	HES	LAU
GARBAGE	**62.1/65.2**	1.2/1.9	27.3/16.8	9.3/16.1
CONSENT	24.2/15.4	**47.3/65.9**	26.4/17.6	2.2/1.1
HESITATION	9.4/13.4	4.0/7.9	**85.6/77.7**	1.0/1.0
LAUGHTER	20.6/15.9	1.6/0.0	14.3/4.8	**63.5/79.4**

Table 4: Confusion Matrix for LSTMs on TUM AVIC using Audio (left, each) and Audio+Shape (right, each) features.

speech recognition, deserves further investigation.

In Table 4 confusions are additionally shown for the audio features and the best fusion case—audio and shape features. One observes that HESITATION is better classified by audio only, while the other classes benefit from the fusion. Apart from the expectable higher number of confusions of any other with the GARBAGE class, more confusions occur between CONSENT and HESITATION, which is explicable by their phonetically partly similar sturcture ("mhm" vs. "hmm").

4. CONCLUSIONS

We introduced a novel audiovisual feature-level fusion by LSTM RNN for the computational assessment of non-linguistic vocalisations in conversational speech. In our experiments, adding shape to audio features improved classification accuracies highly significantly, yet, the further addition of appearance features was observed inferior. A reference approach able to exploit suprasegmental effects by SVM based on statistical functionals fell behind the frame-level modelling with long-term memory. An obvious next step will be the inclusion in an audiovisual speech recognition framework. In addition, further network topologies and bottle-neck feature architectures will be of particular interest.

5. REFERENCES

[1] R.R. Provine, "Laughter punctuates speech: linguistic, social and gender contexts of laughter," *Ethology*, vol. 15, pp. 291–298, 1993.

[2] K. Laskowski, "Contrasting Emotion-Bearing Laughter Types in Multiparticipant Vocal Activity Detection for Meetings," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4765–4768, IEEE.

[3] K.P. Truong and D.A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, pp. 144–158, 2007.

[4] N. Campbell, "Whom we laugh with affects how we laugh," in *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, Jürgen Trouvain and Nick Campbell, Eds., Saarbrücken, 2007, pp. 61–65.

[5] J.-A. Bacharowski and M.J. Smoski, "The acoustic features of human laughter," *J. Ac. Soc. Am.*, vol. 110, no. 3, pp. 1581–1597, 2001.

[6] J. Trouvain, "Segmenting Phonetic Units in Laughter," in *Proc. ICPhS*, Barcelona, 2003, pp. 2793–2796.

[7] S. Petridis and M. Pantic, "Audiovisual Laughter Detection based on Temporal Features," in *Proc. ICMI*, 2008, pp. 37–44.

[8] S. Petridis, A. Asghar, and M. Pantic, "Classifying laughter and speech using audio-visual feature prediction," in *IEEE ICASSP*, 2010, pp. 5254–5257.

[9] A. Ito, Wang Xinyue, M. Suzuki, and S. Makino, "Smile and laughter recognition using speech processing and face recognition from conversation video," in *Intern. Conf. on Cyberworlds, 2005*, 2005, pp. 8–15.

[10] P. Pal, A.N. Iyer, and R.E. Yantorno, "Emotion detection from infant facial expressions and cries," may. 2006, vol. 2.

[11] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," *LNCS*, vol. 5237, pp. 137 – 148, 2008.

[12] S. Escalera, E. Puertas, P. Radeva, and O. Pujol, "Multi-modal laughter recognition in video conversations," pp. 110 –115, 2009.

[13] S. Scherer, F. Schwenker, N. Campbell, and G. Palm, "Multimodal laughter detection in natural discourses," *Human Centered Robot Systems*, pp. 111–120, 2009.

[14] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *Int'l Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 97–104.

[15] D. Gonzalez-Jimenez and J. L. Alba-Castro, "Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry," *IEEE Trans. Inform. Forensics and Security*, vol. 2, no. 3, pp. 413–429, 2007.

[16] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Principal Component Analysis of Image Gradient Orientations for Face Recognition," in *IEEE FG, to appear*, 2011.

[17] W. Ruch and P. Ekman, "The expressive pattern of laughter," *Emotion, qualia, and consciousness*, pp. 426–443, 2001.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Florence, Italy, 2010, ACM.

[19] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, TUM, 2008.

[20] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal*, vol. 27, pp. 1760–1774, 2009.

[21] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. INTERSPEECH 2010*, Makuhari, Japan, 2010, pp. 2794–2797.

[22] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Perception and Interactive Technologies for Speech-based Systems*. 2008, vol. LNCS 5078, pp. 99–110, Springer.