AUDIOVISUAL VOCAL OUTBURST RECOGNITION IN NOISY ACOUSTIC CONDITIONS

Florian Eyben[1]*, Stavros Petridis*[2]*, Björn Schuller*[1]*, Maja Pantic*[2,3]

[1]Institute for Human-Machine-Communication, Technische Universität München, Munich, Germany
[2]Department of Computing, Imperial College London, London, UK
[3]EEMCS, University of Twente, Enschede, Netherlands

eyben@tum.de

ABSTRACT

In this study, we investigate an audiovisual approach for classification of vocal outbursts (non-linguistic vocalisations) in noisy conditions using Long Short-Term Memory (LSTM) Recurrent Neural Networks and Support Vector Machines. Fusion of geometric shape features and acoustic low-level descriptors is performed on the feature level. Three different types of acoustic noise are considered: babble, office and street noise. Experiments are conducted on every noise type to asses the benefit of the fusion in each case. As database for evaluations serves the INTERSPEECH 2010 Paralinguistic Challenge's Audiovisual Interest Corpus of human-to-human natural conversation. The results show that even when training is performed on noise corrupted audio which matches the test conditions the addition of visual features is still beneficial.

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

1. INTRODUCTION

Non-linguistic vocal outbursts are defined as very brief, discrete, nonverbal expressions of affect in both the face and voice [1]. Humans are very good at recognising emotions just by hearing such vocalisations, which suggests that they convey emotion related information. While a growing number of efforts towards automatic recognition of vocal outbursts is recently reported, most of these are based only on audio signals and aimed at automatic laughter recognition [2–4]. Similarly to speech perception by humans, it has been recently demonstrated that laughter is perceived as more audible when the facial expression is visible [5]. Therefore lately, few efforts towards audiovisual recognition of non-linguistic vocal outbursts have been reported including mainly automatic classification of audiovisual laughter episodes [6–8].

In our previous work [9], we performed classification of audiovisual episodes of laughter, consent and hesitation. We compared two different types of features, shape and appearance features, and we found that the combination of audio and shape features leads to an improvement over the audio-only classification approach. Appearance features did not result in an improvement, possibly due to the use of spontaneous data which contain large head movements and non-frontal poses.

In this contribution, we extend our previous work and investigate the performance of the audiovisual classification of vocal outbursts in the presence of noise. A common experiment in the literature, is to train an audiovisual classifier on clean data and then test it on noisy data. As expected, the performance decreases but it remains higher than the performance of the audio-only classifier since the addition of visual features provides extra information which is not corrupted by acoustic noise. However, it is not clear if the improvement observed in noisy conditions is simply because the audio features are degraded, so the addition of clean visual features helps, or due to the complementary information carried by the visual features. In order to investigate this hypothesis, experiments are conducted on matched noisy training and testing condition using 3 different noise types, *babble*, *street* and *office* noise. The results in this study show that the addition of the visual information is beneficial even when the audio classifiers are trained and tested on noisy data under matched conditions, suggesting that the visual features carry indeed complementary information.

The audio and visual modalities are fused at feature-level and classification is performed using Support Vector Machines (SVMs) and Long-Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs). The targets of interest include conversational consent, hesitation and laughter 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: Section 2 describes the audio and visual features used, the experimental protocol and data are discussed in Section 3 and finally results are presented in Section 4.

2. FEATURES

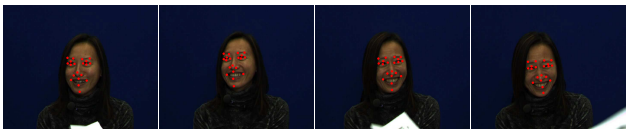
2.1. Audio features

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 geometric shape based set. It is known that the first coefficients suffice for non-linguistic assessment [9].

Acoustic features have been calculated using our open-source extractor openSMILE [10] 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.

2.2. Video features

Given the results presented in [9] we decided to use shape features only, rather than investigating shape and appearance features. Initially, we track 20 facial points using the Patras-Pantic particle filtering tracking scheme [11] as shown in Fig. 1 to 3. For each video



(a) Frame 1 (b) Frame 13 (c) Frame 26 (d) Frame 38

Fig. 1: Example of LAUGHTER from the TUM AVIC corpus.

(a) Frame 1 (b) Frame 7 (c) Frame 14 (d) Frame 22

Fig. 2: Example of HESITATION from the TUM AVIC corpus.

(a) Frame 1 (b) Frame 4 (c) Frame 8 (d) Frame 12

Fig. 3: Example of CONSENT from the TUM AVIC corpus.

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 [12], 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 [12]. This set is referred to as '*Shape*' in the ongoing.

3. EXPERIMENTAL PROTOCOL

3.1. Data

We prepared a data set based on the TUM Audio-Visual Interest Corpus (TUM AVIC). The spoken content, including vocal outbursts (also referred to as non-linguistic vocalisations), is transcribed on the word level. For a detailed description of TUM AVIC we refer to [13]. We follow the official partitioning of the corpus as was used for the INTERSPEECH 2010 Paralinguistic Challenge [14]. 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): GARBAGE (420 / 161), CONSENT (218 / 91), HESITATION (731 / 403), LAUGHTER (204 / 63).

For the experiments described in section 4 each example is corrupted by one of the three noise types considered in this paper: babble noise, street noise, and office noise. For babble and street noise we use the Aurora noise samples (cf. [15]). The office noise consists of typical sounds occurring in a busy office environment, such as typing, printer machines, writing, beep sounds, and occasional talk in the background. The noise samples have been sampled from YouTube^{TM} videos which contained office environment noise recordings.

Our noise samples have the following lengths: one minute for *street* noise, 4 minutes for *babble* noise, and 47 minutes for *office* noise. When overlaying an audio chunk (a single isolated non-verbal) with noise a random region of the noise sample is selected, and scaled accordingly to match the desired Signal-to-Noise-Ratio (SNR) before adding it to the audio chunk. Independent parts of the noise samples are used for training and evaluation sets. The SNRs are computed based on Root Mean Square (RMS) amplitudes $A_{sig}^{(rms)}$ and $A_{noise}^{(rms)}$ of signal and noise chunks, respectively, according to the following equation:

$$SNR = 20 \log_{10} \frac{A_{sig}^{(rms)}}{A_{noise}^{(rms)}} \quad (1)$$

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.

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

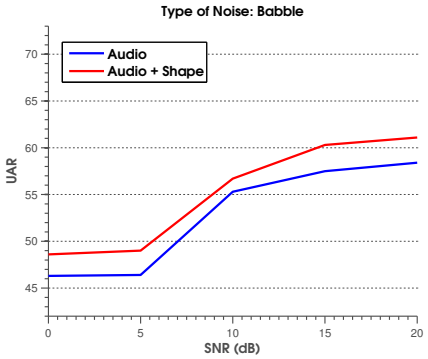
Table 2: Set of 7 functionals used to convert low-level feature contours of variable length to a fixed length vector for static classification with SVM.

3.2. Classification

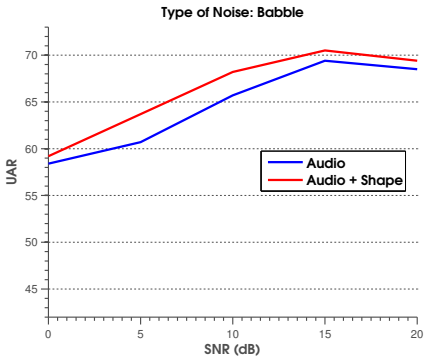
In the experiments presented here we consider isolated non-linguistic vocalisations as in [16]. For all experiments the classifier of choice has been trained on the joint data from the TUM AVIC training and development set, which we will refer to as *training data* in the ongoing. Evaluations have been conducted on the TUM AVIC evaluation set.

We compare the performance of the audio and audiovisual classifiers on matched training and test set noisy conditions on 3 different scenarios: babble noise, street noise and office noise. In other words, both the training and test sets are corrupted by the same type of noise with varying degrees of SNR from 0 dB to 20 dB.

Audio and shape features are extracted at different frame rates, 100 fps and 25 fps, respectively, so shape features are upsampled simply by copying each feature vector 4 times in order to match the audio frame rate. Then the audio and shape features are concatenated for feature-level fusion. 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 SVMs with

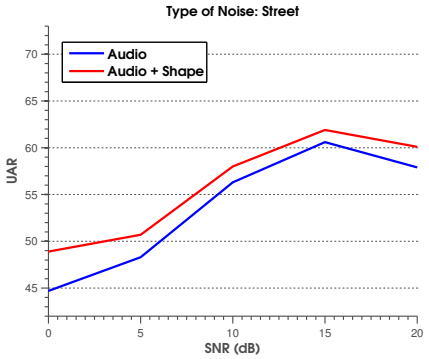


(a) SVM

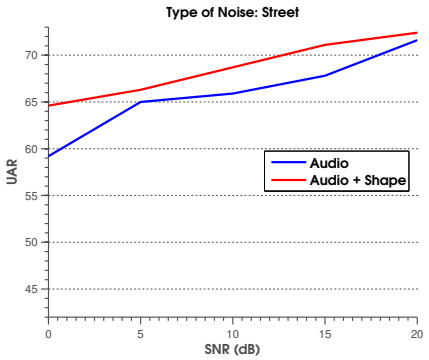


(b) LSTM

Fig. 4: Babble noise.

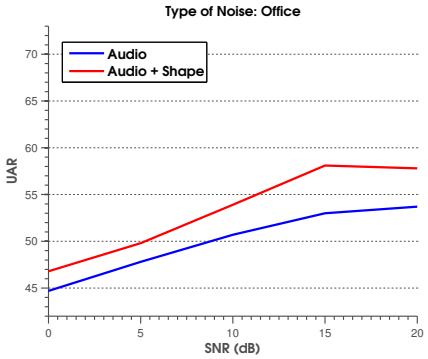


(a) SVM

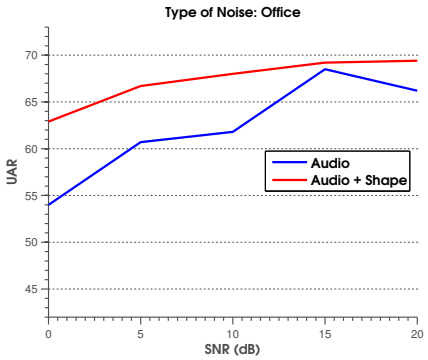


(b) LSTM

Fig. 5: Street noise.



(a) SVM



(b) LSTM

Fig. 6: Office noise.

a radial basis function (RBF) kernel function are employed in a subsequent classification step for reference.

In principle LSTM networks are an extension of recurrent neural networks, equipped with an enhanced memory feature. Although 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 [17]. 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. 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.

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 (one for each class of non-linguistic vocalisations).

For multimodal classification of isolated vocal outbursts, 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. 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.

The static classification approach (by SVM) is similar to the one introduced in [16], 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 descriptors (Audio and Shape features) as used for the proposed LSTM-RNN approach. We then apply a small set of functionals (table 2) to the (fused) set of low-level descriptors' contours.

4. RESULTS

Table 3 shows the unweighted average recall (UAR) rate for the clean data. It is clear that for both types of classifiers the audiovisual fusion outperforms the audio-only approach, and LSTM outperforms SVM, as was already previously shown. Results for the three different noise scenarios are plotted in Figs. 4 to 6. The x-axis in each plot shows the SNR which is used both for training and testing. In all cases, the audiovisual approach (red line) for both classifiers leads to improved performance over the audio-only approach. This indicates that the improvement in performance, in noisy au-

UAR [%]	LSTM	SVM
Audio	67.6	55.3
Shape	41.1	37.3
Audio+Shape	72.3	57.7

Table 3: Results for audiovisual non-linguistic vocalisation classification on TUM AVIC on clean data without the addition of noise. Unweighted average (UAR) of class-wise recall rates are reported. Details in the text.

[%] as →	GAR	CON	HES	LAU
GARBAGE	**55.3/53.3**	2.5/0.4	38.1/36.4	4.1/9.9
CONSENT	16.8/16.8	**39.6/47.3**	40.7/34.1	2.9/1.8
HESITATION	12.7/13.9	3.6/7.9	**82.2/75.5**	1.5/2.7
LAUGHTER	22.7/14.3	1.1/2.1	24.3/10.6	**51.9/73.0**

Table 4: Confusion Matrix for LSTMs on TUM AVIC using Audio (left, each) and Audio+Shape (right, each) features. The confusion matrix is the average over the three types of noise at 0 dB.

dio conditions, achieved by the addition of the visual features is due to their complementary information which can be useful when the audio channel is noisy. If it was simply the result of the degraded performance of the audio features (which is the case when training on clean data and testing on noisy data) when training and testing on noisy conditions, no improvement should have been observed.

It is noteable that the 20 dB SNR case performs slightly below the 15 dB case in some configurations (SVM for street noise, and LSTM for babble and office noise). Interestingly, this effect is opposite for SVM and LSTM regarding the noise type. The 15dB figures are almost at the level of the clean results (cf. table 3). We attribute this to the fact that adding a certain (small) amount of noise to the training data improves the generalisation ability of the classifier, as more variance in the training data is observed. This fact is exploited deliberately when training LSTM-RNN, where often white noise is added to the features of the training partition. In our case the test partition is also corrupted, but we could expect the test results to be superior to the clean case when testing on the clean audio with the classifier trained on the 15 dB SNR noise corrupted versions. These are very interesting issues that were discovered by this contribution. They deserve more attention in follow-up work.

In Table 4 confusions are shown for the audio features and the fusion case – audio and shape features. The results are averaged over the three types of noise at 0 dB. One observes that HESITATION is better classified by audio only, while the other classes benefit from the fusion. A possible explanation is that Hesitation does not involve much movement in the face, compared to consent (nodding) and laughter (periodic movement). Apart from the expectable higher number of confusions of any other with the GARBAGE class, more confusions occur between CONSENT and HESITATION – for the audio only case – which is explicable by their phonetically partly similar structure ("mhm" vs. "hmm"). When adding shape features, the nods which often occur along with consent can be better detected.

5. CONCLUSIONS

We presented an audiovisual feature-level fusion approach by LSTM RNN and SVMs for the computational assessment of non-linguistic vocalisations in conversational speech under noisy conditions. In our experiments, the combination of audio and shape information proved beneficial for all types of noise considered indicating that the information carried by the shape features can be particularly helpful when the audio channel is noisy.

An obvious next step will be the evaluation of noisy conditions for shape feature extraction with clean and corrupted audio, as well as mismatched noisy conditions in order to further investigate the benefits and limitations of the audiovisual fusion. Moreover, the effects of training with noise corrupted data on the recognition performance on clean data needs to be investigated in detail. Together, this might lead to the next great step in improving performance of non-linguistic and paralinguistic classification performance.

6. REFERENCES

[1] K. Scherer, "Affect bursts," in *Emotions: Essays on emotion theory*, S. van Goozen, N. van de Poll, and J. Sergeant, Eds., pp. 161–193. 1994.

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

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

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

[5] T.R. Jordan and L. Abedipour, "The importance of laughing in your face: Influences of visual laughter on auditory laughter perception," *Perception*, vol. 39, no. 9, pp. 1283–1285, 2010.

[6] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Trans. on Multimedia*, vol. 13, no. 2, pp. 216–234, April 2011.

[7] S. Petridis, M. Pantic, and Cohn. J. F., "Prediction-based classification for audiovisual discrimination between laughter and speech," in *Proc. FG 2011*. 2011, IEEE.

[8] 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.

[9] F. Eyben, S. Petridis, B. Schuller, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks," in *Proc. ICASSP 2011, Prague, Czech Republic*, 2011, pp. 5844–5847.

[10] 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.

[11] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *Proc. FG 2004*, 2004, pp. 97–104.

[12] 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.

[13] 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.

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

[15] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP '00*, Beijing, China, October 2000.

[16] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Proc. PIT'08*. 2008, vol. LNCS 5078, pp. 99–110, Springer.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.