The 14th International Conference on Auditory-Visual Speech Processing
25-26 August 2017, Stockholm, Sweden

End-to-End Audiovisual Fusion with LSTMs

Stavros Petridis[1], *Yujiang Wang*[1], *Zuwei Li*[1], *Maja Pantic*[1,2]

[1]Dept. Computing, Imperial College London
[2]EEMCS, University of Twente

stavros.petridis04@imperial.ac.uk, m.pantic@imperial.ac.uk

Abstract

Several end-to-end deep learning approaches have been recently presented which simultaneously extract visual features from the input images and perform visual speech classification. However, research on jointly extracting audio and visual features and performing classification is very limited. In this work, we present an end-to-end audiovisual model based on Bidirectional Long-Short Memory (BLSTM) networks. To the best of our knowledge, this is the first audiovisual fusion model which simultaneously learns to extract features directly from the pixels and spectrograms and perform classification of speech and nonlinguistic vocalisations. The model consists of multiple identical streams, one for each modality, which extract features directly from mouth regions and spectrograms. The temporal dynamics in each stream/modality are modelled by a BLSTM and the fusion of multiple streams/modalities takes place via another BLSTM. An absolute improvement of 1.9% in the mean F1 of 4 nonlingusitic vocalisations over audio-only classification is reported on the AVIC database. At the same time, the proposed end-to-end audiovisual fusion system improves the state-of-the-art performance on the AVIC database leading to a 9.7% absolute increase in the mean F1 measure. We also perform audiovisual speech recognition experiments on the OuluVS2 database using different views of the mouth, frontal to profile. The proposed audiovisual system significantly outperforms the audio-only model for all views considered when the acoustic noise is high.

Index Terms: Audiovisual Fusion, End-to-end Deep Learning, Audiovisual Speech Recognition

1. Introduction

Audiovisual fusion approaches have been successfully applied to various problems like speech recognition, emotion recognition, laughter recognition and biometric applications. The addition of the visual modality is particularly useful in noisy environments where the performance of audio-only classifiers is degraded. As a consequence, the visual information, which is not affected by acoustic noise, can significantly improve the performance of audio-only classifiers in noisy environments.

Recently, several deep learning approaches for audiovisual fusion have been presented. The vast majority of them follow a two step approach where features are first extracted from the audio and visual modalities and then are fed to a classifier. Ngiam et al. [1] applied principal component analysis (PCA) to the mouth region of interest (ROI) and spectrograms and trained a deep autoencoder to extract bottleneck features. The features from the entire utterance were fed to a support vector machine (SVM) ignoring the temporal dynamics of the speech. Hu et al. [2] used a similar approach where PCA was applied to mouth ROIs and spectrograms and a recurrent temporal multimodal restricted Boltzmann machine was trained to extract features which are fed to an SVM. Ninomiya et al. [3] applied PCA to the mouth ROIs and concatenated Mel-Frequency Cepstral Coefficients (MFCCs) and trained a deep autoencoder to extract bottleneck features which were fed to a Hidden Markov Model (HMM) in order to take into account the temporal dynamics. Mroueh et al. [4] used concatenated MFCCs together with scattering coefficients extracted from the mouth ROI in order to train a deep network with a bilinear softmax layer. Takashima et al. [5] used a convolutional neural network to extract bottleneck features from lip images and Mel-maps which were fed to an HMM. It is clear that none of the above works follows an end-to-end architecture.

Few works have been presented very recently which follow an end-to-end approach for visual speech recognition (lipreading). Wand et al. [6] used a fully connected layer followed by two LSTM layers to perform lipreading directly from raw mouth ROIs. Petridis et al. [7] used a deep autoencoder together with an LSTM for end-to-end lipreading from raw pixels. Assael et al. [8] used a CNN with gated recurrent units for end-to-end sentence-level lipreading.

To the best of our knowledge, the only work which performs end-to-end training for audiovisual speech recognition is [9]. An attention mechanism is applied to both the mouth ROIs and MFCCs and the model is trained end-to-end. However, strictly speaking the system is not completely end-to-end since it requires the extraction of MFCCs features first from the audio signal.

In this paper, we extend our previous work [7] and present an end-to-end audiovisual fusion model for speech recognition and nonlinguistic vocalisation classification which jointly learns to extract audio/visual features directly from raw inputs and perform classification. To the best of our knowledge, this is the first end-to-end model which performs audiovisual fusion from raw mouth ROIs and spectrograms. The proposed model consists of multiple identical streams, one per modality, which extract features directly from the raw images and spectrograms. Each stream consists of an encoder which compresses the high dimensional input to a low dimensional representation. The encoding layers in each stream are followed by a BLSTM which models the temporal dynamics. Finally, the information of the different streams/modalities is fused via a BLSTM which also provides a label for each input frame. We perform classification of nonlinguistic vocalisations on AVIC database achieving state-of-the-art performance for audiovisual fusion, with an absolute increase in the mean F1 measure by 9.7%. The proposed system also results in a absolute increase of 1.9% in the mean F1 measure compared to the audio-only model. In addition, we also perform experiments on audiovisual speech recognition using different lip views, from frontal to profile, on OuluVS2. The end-to-end audiovisual fusion outperforms the audio-only model when the noise level is high and results in the same performance when clean audio is used.

10.21437/AVSP.2017-8

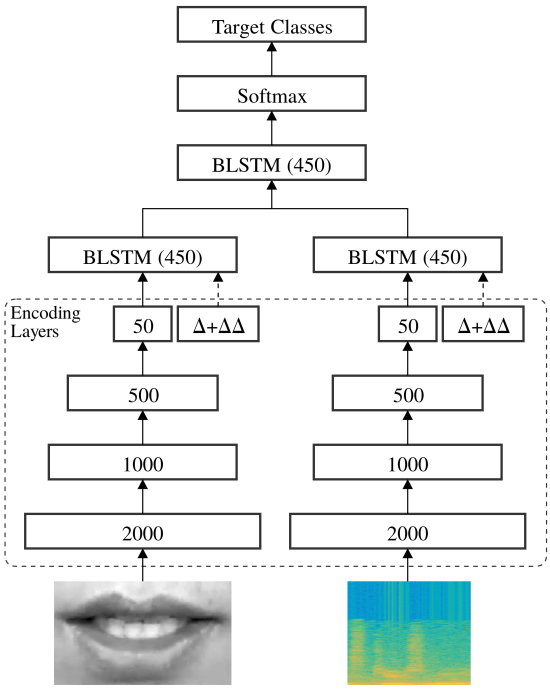

Figure 1: *Overview of the end-to-end audiovisual system. One stream per modality is used for feature extraction directly from the raw images and spectrograms. Each stream consists of an encoder which compresses the high dimensional input image to a low dimensional representation. The Δ and $\Delta\Delta$ features are also computed and appended to the bottleneck layer. The encoding layers in each stream are followed by a BLSTM which models the temporal dynamics. A BLSTM is used to fuse the information from all streams and provides a label for each input frame.*

2. Databases

The databases used in this study are the OuluVS2 [10] and AVIC [11]. The OuluVS2 contains 52 speakers saying 10 utterances, 3 times each, so in total there are 156 examples per utterance. The utterances are the following: "Excuse me", "Goodbye", "Hello", "How are you", "Nice to meet you", "See you", "I am sorry", "Thank you", "Have a good time", "You are welcome". The mouth ROIs are provided and they are downscaled as shown in Table 1 in order to keep the aspect ratio constant. Video is recorded at 30 frames per second (fps) and audio at 48 kHz The unique feature of OuluVS2 is that it provides multiple lip views. To the best of our knowledge it is the only publicly available database with 5 lip views between 0°and 90°.

The AVIC corpus is an audiovisual dataset containing scenario-based dyadic interactions. A subject is interacting with an experimenter who plays the role of a product presenter and leads the subject through a commercial presentation. The subjects role is to listen to the presentation and interact with the experimenter depending on his/her interest on the product.

Annotations for laughter, hesitation, consent and other human noises, which are grouped into one class called garbage, are provided with the database and those are used in this study. In total 21 subjects were recorded, 11 males and 10 females with most subjects being non-native speakers. Similarly to previous works [11, 12, 13] vocalisations that were very short ($\leq$ 120 ms) were excluded. In total, 247, 1136, 308 and 582 examples for the laughter, hesitation, consent and garbage class, respectively, were used. Examples of laughter and hesitation are shown in Fig. 2 and 3, respectively.

Table 1: *Size of mouth ROIs in pixels for each view in the OuluVS2 database.*

Views	0°	30°	45°	60°	90°
Height/Width	29/50	29/44	29/43	35/44	44/30

A video camera was used to record the subject's reaction, positioned in front of him/her, at 25 fps. The audio signal was recorded by a lapel microphone at 44.1 kHz.

AVIC does not provide moouth ROIs so sixty eight points were tracked on the face using the tracker proposed in [14]. The faces were first aligned using a neutral reference frame in order to normalise them for rotation and size differences. This is done using an affine transform using 5 stable points, two eyes corners in each eye and the tip of the nose. Then the center of the mouth is located based on the tracked points and a bounding box with size 90 by 150 is used to extract the mouth ROI. Finally, the mouth ROIs are downscaled to 30 by 50.

3. End-to-end Audiovisual Fusion

The proposed deep learning system for multi-view lipreading is shown in Fig. 1. It consists of two identical streams which extract features directly from the raw input images and the spectrograms [1], respectively. Each stream consists of two parts: an encoder and a BLSTM. The encoder follows a bottleneck architecture in order to compress the high dimensional input image to a low dimensional representation at the bottleneck layer. The same architecture as in [15] is used, with 3 hidden layers of sizes 2000, 1000 and 500, respectively, followed by a linear bottleneck layer. The rectified linear unit is used as the activation function for the hidden layers. The Δ (first derivatives) and $\Delta\Delta$ (second derivatives) [16] features are also computed, based on the bottleneck features, and they are appended to the bottleneck layer. In this way, during training we force the encoding layers to learn compact representations which are discriminative for the task at hand but also produce discriminative Δ and $\Delta\Delta$ features. This is in contrast to the traditional approaches which pre-compute the Δ and $\Delta\Delta$ features at the input level and as a consequence there is no control over their discriminative power.

The second part is a BLSTM layer added on top of the encoding layers in order to model the temporal dynamics of the features in each stream. The BLSTM outputs of each stream are concatenated and fed to another BLSTM in order to fuse the information from all streams. The output layer is a softmax layer which provides a label for each input frame. The majority label over each utterance is used in order to label the entire utterance. The entire system is trained end-to-end which enables the joint learning of features and classifier. In other words, the encoding layers learn to extract features from raw images and spectrograms which are useful for classification using BLSTMs.

[1] Spectrogram frame are computed over a 40 ms windows with 30 ms overlap.

(a) 6532 (b) 6539 (c) 6555 (d) 6569

Figure 2: *Example of laughter from the AVIC corpus, Subject VP4, frames 6532 to 6569.*

(a) 15476 (b) 15479 (c) 15489 (d) 15497

Figure 3: *Example of hesitation from the AVIC corpus, Subject VP8, frames 15476 to 15497.*

Table 2: *Mean F1, Unweighted Average Recall (UAR) and Classification Rates (CR) for the Audio-only classifier (A), Video-only classifier (V) and audiovisual classifiers (A + V) on the AVIC database. Subjects 1 to 7 are used as test set. The highest value in each column is shown in bold.*

Stream	Mean F1	UAR	CR
Current State-of-the-art [13]			
A	54.1	58.7	58.8.0
V	44.0	48.9	48.5
A + V	65.3	64.9	72.6
End-to-End Model			
A	73.1	72.6	79.6
V	46.3	48.4	66.9
A + V	**75.0**	**74.2**	**80.4**

Table 3: *Classification rate of the different views and their combinations with audio on the OuluVS2 database.*

Stream	V	A + V
Frontal	91.8	98.6
30°	87.3	98.7
45°	88.8	98.3
60°	86.4	98.6
Profile	91.2	98.9
Clean Audio	98.5	

4. EXPERIMENTAL SETUP

4.1. Evaluation Protocol

We first partition the data into training, validation and test sets. The same protocol as in [13] is used for the AVIC dataset where the first 7 subjects are used for testing, the next 7 for training and the last 7 for validation.

The protocol suggested in [17] is used for the OuluVS2 dataset where 40 subjects are used for training and validation and 12 for testing. We randomly divided the 40 subjects into 35 and 5 subjects for training and validation purposes, respectively. This means that there are 1050 training utterances, 150 validation utterances and 360 test utterances.

4.2. Preprocessing

Since all the experiments are subject independent we first need to reduce the impact of subject dependent characteristics. This is done by subtracting the mean image, computed over the entire utterance, from each frame.

As mentioned in section 2 the audio and visual features are extracted at different frame rates. Therefore they need to by synchronised. This is achieved by upsampling the visual features, to match the frame rate of the audio features (100fps), by linear interpolation similarly to [18].

Finally, due to randomness in initialisation, every time a deep network is trained the results are slightly different. In order to present a more objective evaluation we run each experiment 10 times and we add the confusion matrices from each run. We use the final confusion matrix in order to compute the performance measures.

4.3. Training

4.4. Single Stream Training

Initialisation: First, each stream is trained independently. The encoding layers are pre-trained in a greedy layer-wise manner using Restricted Boltzmann Machines (RBMs) [19]. Since the input (pixels or spectrograms) is real-valued and the hidden layers are either rectified linear or linear (bottleneck layer) four Gaussian RBMs [19] are used. Each RBM is trained for 20 epochs with a mini-batch size of 100 and L2 regularisation coefficient of 0.0002 using contrastive divergence. The learning rate is fixed to 0.001 as suggested in [19] when visible/hidden units are linear.

As recommended in [19] the data should be z-normalised, i.e. the mean and standard deviation should be equal to 0 and 1 respectively, before training an RBM with linear input units. Hence, each image is z-normalised before pre-training the encoding layers. Similarly, each spectrogram frame is also z-normalised.

End-to-End Training: Once the encoder has been pre-trained then the BLSTM is added on top and its weights are initialised using glorot initialisation [20]. The Adam training algorithm is used for end-to-end training with a mini-batch size of 10 utterances. The default learning rate of 0.001 led to unstable training so it was reduced to 0.0003. Early stopping with a delay of 5 epochs was also used in order to avoid overfitting and gradient clipping was applied to the LSTM layers.

4.5. Audiovisual Training

Initialisation: Once the single streams have been trained then they are used for initialiasing the corresponding streams in the multi-stream architecture. Then another BLSTM is added on top of all streams in order to fuse the single stream outputs. Its weights are initialised using glorot initialisation.

End-to-End Training: Finally, the entire network is trained jointly using Adam with a mini-batch size of 10 utterances. Since the individual streams are already initialised at good val-

ues a lower learning rate is used, 0.0001, to finetune the entire network. Early stopping and gradient clipping were also applied similarly to single stream training.

5. Experiments

In this section we report the results on OuluVS2 and AVIC databases. We have experimented with using the end-to-end audiovisual system shown in Fig. 1 but also with the individual steams, i.e., audio- and video-only classification. In the latter case, we just use the corresponding single stream, encoder + BLSTM.

5.1. Results on AVIC database

Results for the AVIC database are shown in Table 2. Since this is an imbalanced dataset using just the classification rate can be misleading. Hence, we also report the unweighted average recall (UAR) rate and the mean F1 measure over all 4 classes. First of all, we see that the proposed end-to-end system significantly outperforms the current state-of-the-art on the AVIC database. It results in an absolute mean F1 improvement of 19%, 2.3% and 9.7% for the audio-only, video-only and audiovisual classification, respectively.

It is also clear that the audio-only classifier performs much better than the video-only classifier. This is expected since most of the information is carried by the audio channel. In addition, some vocalisations can be accompanied by subtle facial expressions, like hesitation in Fig. 3, or even no facial expression at all. However, the visual modality is still useful and the audiovisual combination using the end-to-end model results in a 1.9% and 1.6% absolute improvement of the mean F1 and UAR, respectively, over the audio-only model.

5.2. Results on OuluVS2 database

We consider a single view scenario where we train and test models on data recorded from a single view. Results are shown in Table 3. This dataset is balanced so we just report the classification rate which is the default performance measure for this database [17]. The best performance in video-only experiments is achieved by the frontal and profile views followed by the 45°, 30°and 60°views. The audio-only model achieves a very high classification accuracy, 98.5%. This is due to the audio signal being clean, without any background noise, and the participants uttering phrases which are much longer than the vocalisations on AVIC database. We also notice that audiovisual fusion leads to negligible improvement over the audio-only model. This is not surprising, given the very high accuracy already achieved by the audio classifier in clean conditions.

In order to test the benefits of audiovisual fusion we have run experiments under varying noise levels. The audio signal is corrupted by additive babble noise from the NOISEX database [21] so as the signal-to-noise ratio (SNR) varies from 0dB to 20dB. Results are shown in Fig. 4. As expected, the audio model is significantly affected by the addition of noise and its performance is degraded more and more as the noise level increases leading to a classification rate of 28.4% at 0dB. All audiovisual models significantly outperform the audio only model due to presence of the visual modality which is not affected by acoustic noise. The classification rate achieved at 0dB varies between 53.3% and 58.1% for 60°and 45°, respectively.

It is worth pointing out, that although there are significant differences in performance between the views in the video-only case, they all result in almost the same performance in the audiovisual case when audio is clean, i.e. no noise added and 15/20dB. However, as the acoustic noise level increases their differences become more evident. It is interesting that the combination of noisy audio with different views does not follow exactly the same pattern as observed in Table 3. Between 0dB and 10dB, the combination of audio with the 45°view is the best one, followed very closely by the combination of audio with the frontal view. The combination of audio and the 60°view is the worst one, which is consistent with Table 3 but surprisingly the combination of audio and profile view is the second worst combination in 0dB. This is an indication that there could be a non-linear interaction between audio and different views when the noisy levels increase but this deserves further investigation.

We should also mention, that beyond 10 dB the performance of the audiovisual fusion model is worse than the performance of the video-only system, which varies between 86.4% (60°view) and 91.8% (frontal view). This is probably due to the fact that the audiovisual system is trained with clean audio data. Given the very high classification accuracy achieved by the audio-only model under clean conditions, the fusion model is probably heavily biased towards audio. The fact that audiovisual fusion results in the same performance as audio-only classification under clean conditions is also an indication towards that direction. As a consequence, when the levels of acoustic noise increase the performance becomes worse than the video-only model, however it is still able to extract some useful information from the visual modality and significantly outperform the audio-only classifier.

Finally, we should also mention that we experimented with CNNs for the encoding layers but this led to worse performance than the proposed system. Chung and Zisserman [22] report that it was not possible to train a CNN on OuluVS2 without the use of external data. Similarly, Saitoh et al. [17] report that they were able to train CNNs on OuluVs2 only after data augmentation was used. This is likely due to the rather small training set. We also experimented with data augmentation which improved the performance but did not exceed the performance of the proposed system.

6. Conclusions

In this work, we present an end-to-end visual audiovisual fusion system which jointly learns to extract features directly from the pixels and spectrograms and perform classification using LSTM networks. Results on audiovisual classification of nonlinguistic vocalisations demonstrate that the proposed model achieves state-of-the-art performance on the AVIC database. In addition, audiovisual speech recognition experiments using different lip views on OuluVS2 demonstrate that the proposed end-to-end model outperforms the audio-only classifier for high level of acoustic noise. The model can be easily extended to multiple streams so we are planning to perform audiovisual multi-view speech recognition and investigate the influence of audio to the different views.

7. References

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. of ICML*, 2011, pp. 689–696.

[2] D. Hu, X. Li *et al.*, "Temporal multimodal learning in audiovisual speech recognition," in *IEEE CVPR*, 2016, pp. 3574–3582.

[3] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, "Integration of deep bottleneck features for audio-visual speech

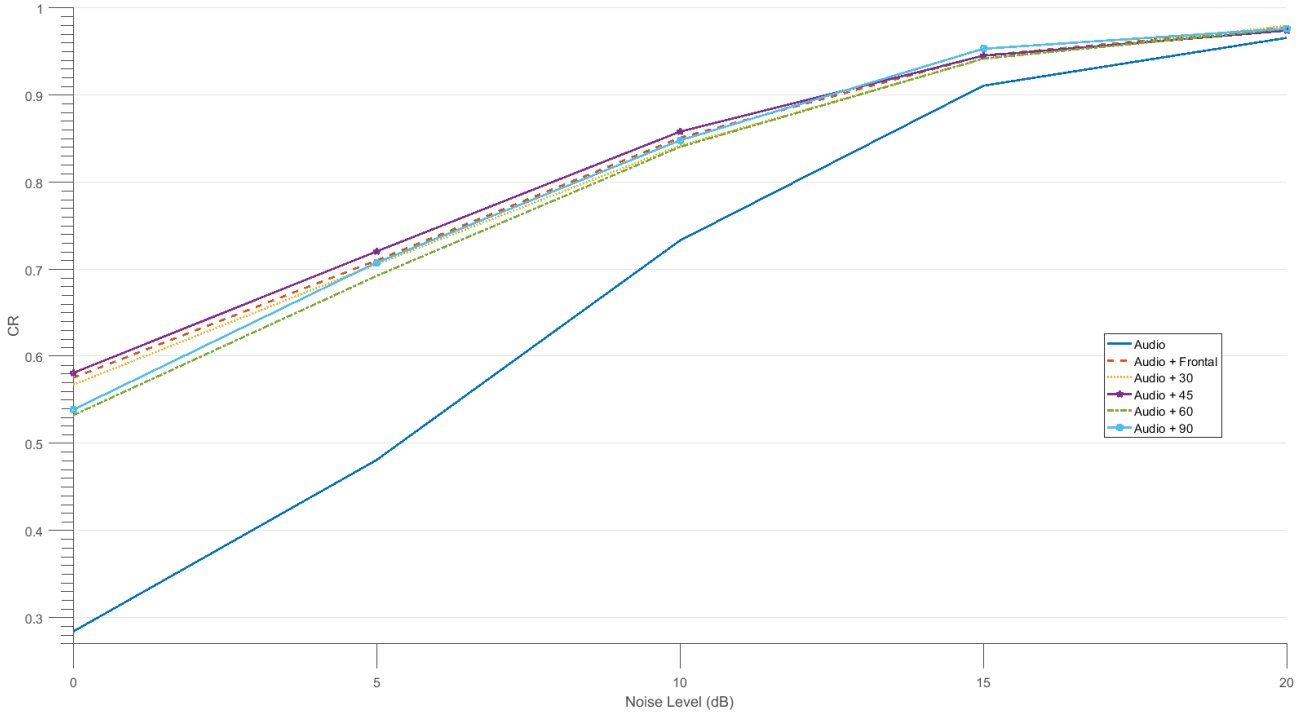

Figure 4: *Classification Rate on the OuluVS2 database for different noise levels. In order to reduce clutter, the video-only performances of the 5 single-view classifiers are omitted. They can be found in Table 3 Best seen in colour.*

recognition," in *Conf. of the International Speech Communication Association*, 2015.

[4] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *IEEE ICASSP*, 2015, pp. 2130–2134.

[5] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, and K. Nakazono, "Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss," *Interspeech 2016*, pp. 277–281, 2016.

[6] M. Wand, J. Koutn, and J. Schmidhuber, "Lipreading with long short-term memory," in *IEEE ICASSP*, 2016, pp. 6115–6119.

[7] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with lstms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017.

[8] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *arXiv preprint arXiv:1611.05358*, 2016.

[10] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *IEEE FG*, 2015, pp. 1–5.

[11] B. Schuller, R. Mueller, F. Eyben, J. Gast, B. Hoernler, M. Woellmer, G. Rigoll, A. Hoethker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Comp.*, vol. 27, no. 12, pp. 1760–1774, 2009.

[12] 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. IEEE ICASSP*, 2011, pp. 5844–5847.

[13] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, 2016.

[14] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE CVPR*, 2014, pp. 1867–1874.

[15] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," vol. 3, p. 175, 2002.

[17] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based cnn for visual speech recognition," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 277–289.

[18] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[19] G. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 599–619.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256.

[21] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[22] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 87–103.