INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Fusion Architectures for Word-based Audiovisual Speech Recognition

Michael Wand, Jürgen Schmidhuber

Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA),
USI & SUPSI, Manno-Lugano, Switzerland
michael@idsia.ch, juergen@idsia.ch

Abstract

In this study we investigate architectures for modality fusion in audiovisual speech recognition, where one aims to alleviate the adverse effect of acoustic noise on the speech recognition accuracy by using video images of the speaker's face as an additional modality. Starting from an established neural network fusion system, we substantially improve the recognition accuracy by taking single-modality losses into account: late fusion (at the output logits level) is substantially more robust than the baseline, in particular for unseen acoustic noise, at the expense of having to determine the optimal weighting of the input streams. The latter requirement can be removed by making the fusion itself a trainable part of the network.

Index Terms: Audio-visual Speech Recognition, Deep Neural Networks, Multimodality, Lipreading, Modality Fusion

1. Introduction

Audiovisual speech recognition [1, 2] has recently been under intensive research, not only because the visual modality can help to compensate for a noisy acoustic data stream, but also as a testbed for systematically experimenting with modality fusion in machine learning. In this paper we follow up on our prior work [3] and systematically investigate fusion of noisy audio and clean video in a word-based speech recognition task, using the established GRID audiovisual corpus [4]. We present and evaluate different neural network fusion architectures, being particularly interested in how our system performs on *untrained* acoustic noises.

The focus of this work is how to merge the acoustic and visual input streams at the architectural level. In the case of HMM-based audiovisual fusion, adaptive weighting of input modalities [5, 6] has been shown to be important for obtaining optimal results; the architectures to obtain this weighting can occasionally become quite complex [7]. Fusion of noisy audio and video has also been performed with neural network architectures [8, 3, 9]. We opt for a classical LSTM-based architecture [10, 11] and present an end-to-end architecture which gives state-of-the-art results on fusing video data with both trained and untrained acoustic noise.

2. Related Work

Audiovisual speech recognition was originally proposed by Petajan [1, 2]; Chiou and Hwang [12] were the first to consider *lipreading*, i.e. purely visual speech recognition, as a standalone task. Since then, a variety of systems were presented, including neural network architectures for lipreading [13, 14, 15, 16, 17, 18] and audiovisual fusion architectures based on HMMs [7] and neural networks [19, 8, 3, 9]. These works are part of the large-scale research effort to develop and validate multimodal fusion architectures, both for classical data-centered machine learning tasks (like audiovisual fusion, emotion recognition [20], multimodal representation learning [21] etc.) and for human-centered systems, which often benefit from multimodality [22]. In the important field of physiology and medicine, a variety of multimodal systems have been proposed [23, 24, 25], see [26, 27] for more examples.

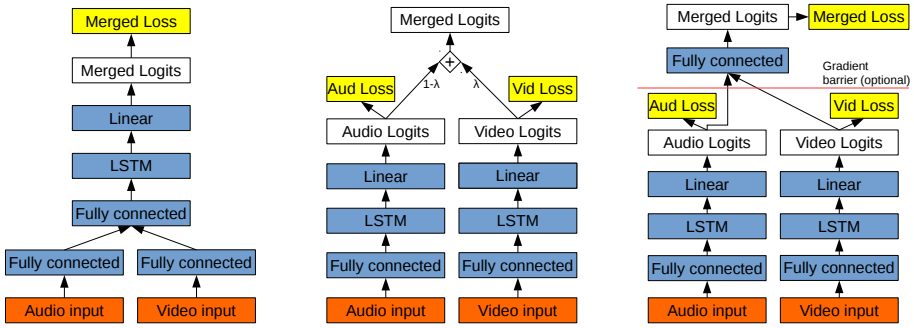
3. Data Corpus

The GRID audiovisual corpus [4] consists of recordings of 34 persons each speaking 1000 sentences, for a total of 28 hours. All sentences follow the pattern *command(4) + color(4) + preposition(4) + letter(25) + digit(10) + adverb(4)*, for example "Place red at J 2, please", where the number in parentheses indicates the possible alternatives, for a total of 51 words. The corpus is constructed so that the probability of each word is independent of its neighboring words. We use the provided word-level segmentation, obtaining our dataset of 6000 single words per speaker. Note that all experiments presented in this study are *speaker-dependent*. The data of each speaker is divided into training, validation, and test set, exactly as in our prior works [14, 28, 18]; validation and test set each contain five examples per word. The validation set is used for early stopping (see section 4), all results reported in this paper are on the test sets. The entire dataset is subdivided into the *development* speakers #1 – #20 and the *evaluation* speakers #22 - #34, speaker 21 is excluded because no video data is available.

Raw **audio** is augmented with noises from the freesound database [29], in particular, we use babble, music, and white noise at {-5dB, 0dB, 5dB} SNR. 27-dimensional log Mel scale features are computed from the raw audio, using a window length of 20 ms and a window shift of 10 ms. Audio feature computation is performed with the OpenSMILE toolkit [30], for superimposing noise to raw audio, we use the open-source acoustic simulator from [31]. From the **video** data, the mouth ROI is extracted with the DLib facial landmark detector [32], using the implementation from [33]. The detected mouth area (landmarks #49 - #68) is enlarged by 10 pixels to the left and to the right, resized to 80x40 pixels, and converted into grayscale for faster processing. For **fusing** the modalities, the video stream is upsampled by a factor of 4 to obtain 100 frames/second as for the audio data. Figure 1 shows an example frame of the corpus, with highlighted mouth area.

Figure 1: *Example frame from GRID, mouth area indicated*

http://dx.doi.org/10.21437/Interspeech.2020-2117



(a) *Intermediate fusion [3]* (b) *Fixed-weight fusion* (c) *Trainable fusion*

Figure 2: *Fusion architectures: Intermediate fusion (baseline), late fusion with fixed stream weights, trainable fusion. The latter comes in two different flavors: The single-modality part and the fusion part can be trained separately, without gradient flow between the parts ("two-step trainable fusion"), or the architecture can be trained end-to-end, averaging all losses ("end-to-end regularized fusion"). "Fully connected" always stands for an arbitrary number of fully connected layers. See text for details.*

4. Architectures and Training

In this study, we consider four different architectures for word-based speech recognition, as follows:

- *Intermediate fusion* (figure 2a) is our baseline from [3]. The network comprises two single-modality parts and a fused part, followed by the final LSTM layer and a softmax loss (the name indicates that modality fusion occurs at an intermediate step in network processing). The whole architecture is trained end-to-end.
- *Fixed-weight fusion* (figure 2b) is a standard late fusion architecture: Two networks, one for audio-based recognition and one for video-based recognition, are trained separately. Fusion is performed by computing the weighted sum of the output logits (before applying the softmax scaling), with a configurable weight (λ for video, $1 - \lambda$ for audio). As we will see in section 5, correctly determining the fusion weight is of utmost importance and can only be done by actually evaluating the fusion system on a suitable data set.

Our main contribution, *Trainable Fusion* (figure 2c) removes the need to choose stream weights; it comes in two variants:

- *Two-step trainable fusion* consists in first training two single-modality recognizers, exactly as for fixed-weight fusion. In a second step, a network is trained to perform joint audiovisual word classification, based on the concatenated output logits of the single-modality networks, whose weights are frozen in this step.
- *End-to-end regularized fusion* uses the same architecture as two-step trainable fusion, but performs only a single training pass, and gradients are simultaneously backpropagated from all three losses through the entire network. The losses may optionally be weighted (but we obtained good results by assigning identical weights to all of them). This architecture is not unlike the intermediate fusion baseline, the difference lies in the side losses which can be interpreted as a kind of regularizer.

In order to allow comparing architectures, we aim to use similar network topologies. In all cases, the audio or video input streams are processed by a series of feedforward fully-connected layers, each followed by the tanh nonlinearity and 50% dropout [14, 28]. In the case of intermediate fusion, the outputs of these sub-networks are concatenated and passed through a further block of fully-connected layers, tanh, and dropout, after which an LSTM layer integrates the sequence information. In all other cases, the LSTM layer is applied on top of the *single-modality* sub-networks. Since we perform word-based recognition, the LSTM output is masked so that only the last frame is considered in both training and backpropagation; this output is immediately passed through a linear layer with 51 neurons, corresponding to each of the words that can be recognized. For fixed-weight fusion and trainable fusion, the output logits of this last layer are further processed, see figure 2.

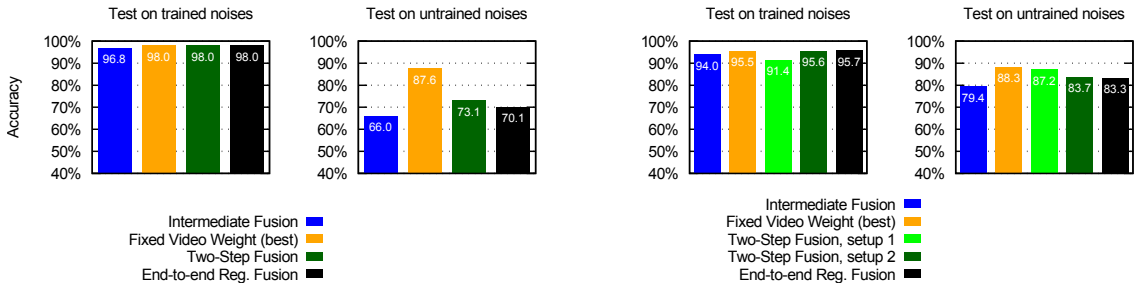
We base our systems on the best architecture from [3]. Our system for intermediate fusion uses two fully connected layers with 128 neurons for each single-modality subnetwork, the fusion part consists of a single fully connected layer and the LSTM, always with 128 neurons/cells. In all other architectures, the *single-modality* parts consist of two fully connected layers and the LSTM, again each with 128 neurons/cells. We keep this part of the architecture fixed since varying the number and size of the network layers within reasonable limits does not have a major effect on the recognition performance [3]. We allow more variation in the fusion parts, see section 5.

We train our systems with a batch size of 64 word samples and the Adam optimizer [34], using the standard learning rate of 0.001 except for two-step trainable fusion, where we train the fusion part (i.e. the second step) with a learning rate of 0.0005. We always perform early stopping (on the validation set of each speaker) with a patience on 30 epochs.

5. Experiments and Results

In this section, we present all our experiments and results, for a variety of training and test data combinations. In particular, we consider systems which are trained on clean audio, and systems where clean audio and audio with injected white noise at three SNR are used for training. *All* systems also use the video stream, for simplicity we do not mention this further. We never use audio with music noise or babble noise for training.

Systems are evaluated on all available audio types (clean audio and audio with white noise, babble noise, and music noise, again at three SNR). We usually report results on *trained* and *untrained* noises: For example, when a system is trained on clean audio, the only trained noise is clean audio, while white noise, babble noise, and music noise at any SNR are untrained.



(a) *Trained on video and clean audio*

(b) *Trained on video and clean + white noise audio*

Figure 3: *Overview of recognition accuracies for different combinations of training and testing data (development speakers)*

When a system is trained on clean audio and white noise audio, trained noises are likewise clean audio and white noise audio, whereas babble noise and music noise are untrained.

5.1. Baseline System and Fixed-weight Fusion

We train the baseline intermediate fusion system on both clean audio, and on clean and white-noise audio. The leftmost (blue-colored) bars of figure 3a and 3b show accuracies for these two cases, averaged over the 20 development speakers: When training on clean audio, accuracies are 96.8% on the trained clean audio and 66.0% on untrained noisy audio, respectively. When training on clean audio and white noise, the recognition accuracy is 94.0% on trained noises and 79.4% on the untrained noises. The accuracy on clean audio is 95.8%. It is immediately clear that the intermediate-fusion system does not perform well on untrained noises.

The second (orange-colored) bars in figures 3a and 3b show results for the fixed-weight fusion system, where we varied the weighting of the video stream between 0.0 and 1.0 in steps of 0.1 and took the *best* average result for each possible combination of training and test noise. We obtain clear improvements in all setups, but in particular when testing on unknown noises: For the system trained on clean audio only, the accuracy improves from 66.0% to 87.6%, for the system trained on clean and white noise audio, the accuracy on the unknown noises improves from 79.4% to 88.3%.

This improvement, however, comes at a price: one needs to determine the correct weighting of the audio and video streams. The importance of this step can be seen from figure 4, where we show the recognition accuracy for diverse combinations of training and test data, plotted versus the video stream weight: For example, when training on clean audio, the best accuracy for testing on clean audio is reached at a video stream weight of 0.5, whereas for unknown audio noises one should choose a video stream weight of 0.7. In the latter case, taking a weight of 0.5 would cause the accuracy to decrease from 87.6% to 83.5%, an error increase of more than 30% relative. When the system is trained on clean audio and white-noise audio, the discrepancy for different video stream weights is less striking, but still present: the optimal video stream weight is 0.4, 0.5, and 0.6 for clean audio, trained noises, and untrained noises, respectively.

Finally, we note that taking video stream weights of 0.0 and 1.0 gives us baseline results for single-modality systems: In particular, the video-only recognition accuracy is 81.4%, the accuracy for training and testing on clean audio is 95.0%. On unknown noises, the audio-only accuracy is between 50% and 60%, and it is clear that both intermediate fusion and fixed-weight fusion substantially improve over the audio-only system.

5.2. Trainable Fusion

One major goal in contemporary machine learning is the development of systems which work well with as little manual intervention as possible. Unfortunately, we see that fixed-weight modality fusion fails in this regard, since it is necessary to manually fix the stream weighting, which depends on the input data. (We assume that the optimal stream weighting varies even more when also the video stream is distorted.)

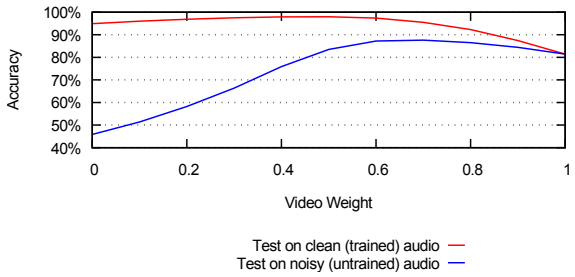
Therefore, it is a natural step to *train* the fusion part of our architecture, based on the existing single-modality networks. This idea leads to the two-step training method described in section 4. We obtained best results with a network consisting of two feedforward layers (with tanh nonlinearity and dropout after the first of them) with 256 neurons each. We also reduced the Adam learning rate to 0.0005: at an early stopping patience of 30 epochs, this helped us to reduce overfitting.

Results are displayed as green-colored bars in figures 3a and 3b. In the case of clean-audio training, the resulting accuracy of 98.0% on clean-audio test data is comparable to fixed-weight fusion, however on noisy audio, the result is substantially worse (yet still much better than the baseline). When we allow both clean audio and white-noise audio for training, we have two options: we can pretrain the audio-only classifier on clean data only (setup 1), or we can use both clean data and white-noise audio (setup 2). In both cases, we use clean audio and white-noise audio for training the fusion part of the system.

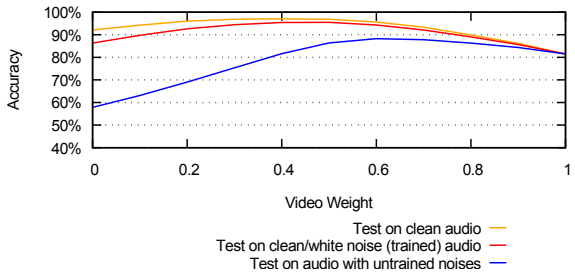
These setups behave strikingly different: setup 2 is much better on trained noises (95.6% vs 91.4% accuracy), setup 1 performs much better on untrained noises (87.2% vs 83.7% accuracy); the accuracy of 87.2% is almost as good as with fixed-weight fusion, *without* requiring access to oracle data for determining the optimal stream weight. In particular, if we set the video stream weight to 0.5, which is optimal on *known* noises, the accuracy of fixed-weight fusion on *unknown* noises drops to 86.4%, slightly worse than with two-step trainable fusion.

We thus have established that we can train the fusion part of our neural network and obtain state-of-the-art results, both for known acoustic noise, and for unknown acoustic noises (underlying babble and music) which are substantially different from the artificial white noise which we used to train the system. We finally cast this method into an end-to-end trainable architecture by training the whole architecture in a single step, with a joint loss computed as the average of the three single losses (see figure 2c, without the gradient barrier).

We ran an architecture search on the *fusion* part of this



(a) *Trained on video and clean audio*



(b) *Trained on video and clean + white noise audio*

Figure 4: *Recognition accuracies of fixed-weight fusion, for different combinations of training and testing data (development speakers)*

system, resulting in the following topologies. For clean-audio training, the fusion part of the network comprises two layers with 128 neurons each, when training on clean audio and white-noise audio, we obtained slightly better results by reducing the size of the two fusion layers to 64 neurons each. The single-modality parts of the network retain the standard topology (two feedforward layers with 128 neurons each, followed by a 128-cell LSTM layer).

Results are given as black bars in figures 3a and 3b. We obtain roughly identical results as with the two-step fusion training process, with one unfortunate exception: setup 1 (where the parts use *different* training data) cannot directly be emulated with end-to-end regularized fusion. (Clearly, one could apply a conditional gradient barrier, creating a hybrid between two-step and end-to-end regularized fusion, but this is complicated and does not offer any actual benefit.) We finally remark that the accuracy substantially degrades when the weight of the side losses is set to zero. This shows that the power of end-to-end regularized fusion stems not only from the bottleneck-style constriction at the single-modality logits layers, but also from actually injecting the side losses.

5.3. Evaluation

Based on the above experiments, we formulate hypotheses to be investigated on the held-out evaluation speakers:

1. All novel fusion methods presented in this paper (fixed-weight fusion, setup 2 of two-step trainable fusion, and end-to-end regularized fusion) improve upon the baseline, for both trained and untrained noise.
2. When training the two-step fusion system on clean audio data and white-noise audio data, *setup 1* improves over *setup 2* in the case of untrained noise.

We will perform statistical validation (one-sided t-test on the data series of the 13 evaluation speakers) for both these claims. Finally, we evaluate whether any of the trainable fusion architectures can fully substitute fixed fusion (with the manually determined optimal stream weighting) in all relevant cases.

Table 1 summarizes results on the evaluation set and gives the p-values relevant to evaluate claim 1, i.e. for the improvement over the baseline. It can be seen that all improvements are significant ($p < 0.05$). When training on clean audio and white noise, the improvement of setup 1 over setup 2 in the case of untrained noises is likewise significant ($p = 0.036$).

Finally, consider the table in its entirety. We see that whenever we test a system on trained noises, two-step trainable fusion (and to some extent end-to-end regularized fusion) can fully substitute fixed-weight fusion, thus we have reached our goal on improving over the original baseline without requiring a manual selection of the modality weights. In the case of untrained noises, the result is slightly less promising: While fixed-weight fusion reaches an 88.2% accuracy, the best competing system (two-step trainable fusion, setup 1) achieves only 86.0% accuracy. This is however still much better than the baseline and demonstrates that training our system in a versatile manner, using different noises for both stages, greatly improves over a more naïve approach.

6. Conclusions

In this study, we have presented network topologies and training strategies for audiovisual fusion in a noisy-audio speech recognition task, showing significant improvement over the baseline. Comparing the novel architectures with the intermediate-fusion baseline system, we see that training for single-modality recognition acts as a kind of regularizer, causing a substantially higher accuracy in particular on *untrained* acoustic noises. We believe that exploring this regularization viewpoint could shed more light on open questions in the field of modality fusion, particularly when the underlying data contains domain variation as in our study. Clever usage of data from different (noise) domains during training the multimodal classifier is of great benefit—this key result should generalize well beyond the specific field of audiovisual speech recognition, allowing to improve a variety of systems in which multiple modalities with different quality parameters contribute towards a common task.

Trained on clean audio				
	Tested on			
	trained noise		untrained noise	
	Acc.	p-value	Acc.	p-value
Intermediate Fusion	97.3%		63.1%	
Fixed-weight Fusion	97.9%	0.034	85.8%	<0.0001
Two-step train. Fusion	98.0%	0.006	69.3%	<0.0001
End-to-end reg. Fusion	97.9%	0.016	66.9%	<0.0001
Trained on clean audio and white-noise audio				
	Tested on			
	trained noise		untrained noise	
	Acc.	p-value	Acc.	p-value
Intermediate Fusion	93.8%		81.3%	
Fixed-weight Fusion	95.1%	0.002	88.2%	<0.0001
Two-step train. Fusion (setup 1)	90.5%		86.0%	
Two-step train. Fusion (setup 2)	95.1%	0.001	84.8%	<0.0001
End-to-end reg. Fusion	94.7%	0.001	83.7%	<0.0001

Table 1: *Recognition accuracies and p-values (improvement over intermediate fusion) for different systems and setup, on the evaluation speakers. All results are significant.*

7. References

[1] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition," in *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, 1984.

[2] ——, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1984.

[3] M. Wand, N. T. Vu, and J. Schmidhuber, "Investigations on End-to-End Audiovisual Fusion," in *Proc. ICASSP*, 2018, pp. 3041 – 3045.

[4] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421 – 2424, 2006.

[5] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning Dynamic Stream Weights For Coupled-HMM-Based Audio-Visual Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 863 – 876, 2015.

[6] S. Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa, "Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR," in *Proc. Interspeech*, 2016, pp. 2135 – 2139.

[7] A. H. Abdelaziz, "Comparing Fusion Models for DNN-Based Audiovisual Continuous Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 475 – 484, 2017.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," in *Proc. ICML*, 2011, pp. 689 – 696.

[9] G. Sterpu, C. Saam, and N. Harte, "Attention-based Audio-Visual Fusion for Robust Automatic Speech Recognition," in *Proc. ICMI*, 2018, pp. 111 – 115.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735 – 1780, 1997.

[11] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[12] G. I. Chiou and J.-N. Hwang, "Lipreading from Color Video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192 – 1195, 1997.

[13] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using Convolutional Neural Network," in *Proc. Interspeech*, 2014, pp. 1149 – 1153.

[14] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," in *Proc. ICASSP*, 2016, pp. 6115 – 6119.

[15] S. Petridis and M. Pantic, "Deep Complementary Bottleneck Features for Visual Speech Recognition," in *Proc. ICASSP*, 2016, pp. 2304 – 2308.

[16] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in *Proc. ACCV*, 2016.

[17] G. Sterpu, C. Saam, and N. Harte, "Can DNNs Learn to Lipread Full Sentences?" arXiv:1805.11685, 2018.

[18] M. Riva, M. Wand, and J. Schmidhuber, "Motion Dynamics Improve Speaker-independent Lipreading," in *Proc. ICASSP*, 2020, pp. 4407 – 4411.

[19] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Communications Magazine*, pp. 65 – 71, 1989.

[20] E. Ghaleb, M. Popa, E. Hortal, and S. Asteriadis, "Multimodal fusion based on information gain for emotion recognition in the wild," in *Proc. IntelliSys*, 2017, pp. 814 – 823.

[21] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A Deep Visual-Semantic Embedding Model," in *Proc. NIPS*, 2013.

[22] M. Turk, "Multimodal Interaction: A Review," *Pattern Recognition Letters*, vol. 36, pp. 189 – 195, 2014.

[23] X. Lei, P. A. Valdes-Sosa, and D. Yao, "EEG/fMRI fusion based on independent component analysis: integration of data-driven and model-driven methods," *Journal of Integrative Neuroscience*, vol. 11, no. 3, pp. 313 – 317, 2012.

[24] A. V. de Vel, K. Cuppens, B. Bonroy, M. Milosevic, K. Jansen, S. V. Huffel, B. Vanrumste, L. Lagae, and B. Ceulemans, "Non-EEG seizure detection systems and potential SUDEP prevention: State of the art," *Seizure*, vol. 22, pp. 345 – 355, 2013.

[25] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-Based Spoken Communication: A Survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257 – 2271, 2017.

[26] D. Lahat, T. Adalı, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449 – 1477, 2015.

[27] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017.

[28] M. Wand and J. Schmidhuber, "Improving Speaker-Independent Lipreading with Domain-Adversarial Training," in *Proc. Interspeech*, 2017, pp. 3662 – 3666.

[29] F. Font, G. Roma, and X. Serra, "Freesound Technical Demo," in *Proc. ACMMM*, 2013, pp. 411 – 412.

[30] F. Eyben, F. Weninger, F. Groß, and p. . Björn Schuller.In Proceedings of the 21st ACM international conference on Multimedia, 2013, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in *Proc. ACMMM*, 2013, pp. 835 – 838.

[31] M. Ferràs, S. Madikeri, P. Motlicek, S. Dey, and H. Bourlard, "A Large-Scale Open-Source Acoustic Simulator for Speaker Recognition," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 527 – 531, 2016.

[32] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755 – 1758, 2009.

[33] A. Rosebrock, "https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python."

[34] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.